Modelling studies of coiled-coil protein in wool fibre

A thesis submitted in partial fulfilment
of the requirements for the Degree
of
Doctor of Philosophy in Chemistry
at the
University of Canterbury
by

Kaiwan Gan

University of Canterbury, Christchurch, New Zealand

November 1995

To my respected parents,
my loving wife, and my dearest friends.

Abstract

A multiple regression analysis has established a non-linear relationship between the backbone dihedral angles and the C^{α} coordinates obtained from the X-ray crystal structures of fourteen proteins. The regression equations have been applied to predict specific dihedral angles of each residue in the backbone of twenty-four proteins. Overall this method (NLRDT) predicts values of ϕ and ψ within a $\pm$ 45° window of those found in the X-ray structure with an accuracy of 94% and 91% and within a $\pm$ 30° window of 88% and 81%.

Two methods for the assignment of motif from C^{α} coordinates are reported. For the first method motif is assigned from the dihedral angles predicted using the regression equations. If the predicted dihedral angles of a residue fall in the range of $-15° > \phi > -90°$ and $-10° > \psi > -70°$, the residue is assigned as in an α-helix; and in the range of $-90° > \phi > -150°$ and $95° < \psi < 170°$ as in a β-sheet. By the second method motif of the *i*th residue is assigned from the distance C_{i-1}^{α} to C_{i+2}^{α} (v_6) and torsional angle C_{i-1}^{α}, C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α} (v_{13}). If these values for a residue fall in the range $v_6 <$ 6.0 Å and $100° > v_{13} > 0°$ the residue is assigned as in an α-helix. If the values are in a range $v_6 > 8.7$ Å and $|v_{13}| > 100°$ the residue is assigned as in a β-sheet. For the twenty four proteins 23.7% of the residues by the former method and 19.6% by the latter method are assigned differently from the PDB.

A Monte Carlo Protein Building (MCPB) method to construct the backbone and C^{β} atomic coordinates of twenty-four proteins from known C^{α} coordinates is reported. The method selects values of dihedral angles from either $\pm$ 30° windows of the dihedral angle calculated for that amino acid by the Non Linear Regression Distance Torsion (NLRDT) method or from ranges established from a statistical analysis of the relationship between dihedral angles of the backbone and C^{α} coordinates for a protein data base. The averaged coordinates from ten backbone models of a protein were used to define a mean structure that was refined by energy minimisation using the AMBER force field (GB/SA). By the latter method the average atomic deviation and r.m.s.d. of

the backbone and C^{β} atoms is between 0.14 Å and 0.32 Å (average 0.22 Å) and 0.22 Å and 0.61 Å (average 0.43 Å) respectively. A comparison with other methods is made.

A model of nine proteins including side chain atoms have been built from the known C^{α} coordinates and amino acid sequences using a Monte Carlo Protein Building Annealing (MCPBA) method. The Cartesian coordinates for the side chain atoms were established with bond lengths and angles selected randomly from within ranges of values previously determined by analysis of fourteen protein crystal structures and with torsional angles randomly selected from -180° to 180°. A simulated annealing technique is used to generate some 300 structures with differing side chain conformations. The atomic coordinates of the backbone atoms are fixed during the simulated annealing process. The coordinates of the side-chain atoms of the 300 low energy conformations are averaged to obtain a mean structure which is minimisation with the C^{α} atoms constrained to their position in the X-ray structure using the OPLS/AMBER force field with the GB/SA water model. The r.m.s.d of the main-chain atoms (without C^{β}) compared with the corresponding crystal structures is in the range 0.20 Å to 0.64 Å with a average value of 0.45 Å. The r.m.s.d of the side-chain atoms is from 1.72 Å to 2.71 Å with an average of 2.26 Å. The r.m.s.d of all atoms is from 1.19 Å to 1.99 Å with an average of 1.61 Å. The method is insensitive to random errors in the C^{α} positions and the computational requirement is modest.

A full atomic model of 7c and 8c-1 coiled coil rod domain in wool protein has been established from the amino acid sequences using the MCPB/MCPBA method. For the particular knob-hole heptad repeat, for the single α-helix the rise per residue is 1.464 Å; the twist angle per residue 102.999°; the number of residues per turn is 3.524 and the pitch 5.171 Å. For the four coiled coil helical segments of the rod domain the pitch is in the range 124 Å to 192 Å (average 172 Å); the radius of the coiled coil varies between 5.24 Å to 5.92 Å; the average value of the radius is 5.56 Å; the average crossing angle of the helices in the coiled coil is 23.0°; the number of residues per major turn is 127.3 and there are 36.2 minor turns in a major turn. The interaction energy between the two α-helical chains in the coiled coil structure is evaluated from van der Waals non-bonded interactions, electrostatic and hydrogen bonding interactions. The

optimum relationship of the α-helical chains to each other established the heptad repeat interaction; 34% of the leucine residues are located at the *d* position. Of the backbone hydrogen bonds in the protein α-helix between residues four apart, 18% have a distance between a donor NH nitrogen and acceptor carbonyl oxygen greater than 3.5 Å .The hydrogen bonds between the side chains of the two α-helices in the coiled coil structure are largely between Arg and Glu, Arg and Asp and Glu and Asp. The distances of the $C\beta$ atoms of cysteine residues are > 4.5 Å. This distance is outside that required for formation of disulphide bonds. The interaction of charged residues with apolar, polar and charged residues in the *a-a, a-d, d-d,* and *d-a* heptad positions accounts for 70% of the interaction energy.

Contents

Chapter One

Introduction

Keratin is a proteinaceous material. Keratin-containing tissues are typically unreactive toward the natural environment and are mechanically strong and durable. It has long been recognised that keratin is not a single material but a complex mixture of sulfur-containing proteins stabilised by disulfide linkages. No precise definition of keratin exists and the term is used simply to denote an insoluble complex of sulfur-containing epidermal proteins. A distinction has been drawn between the so-called soft keratin found in stratum corneum, corns, callouses, and the eponychium of nails, and the hard keratin found in wool, hair, nails, claws, beaks, horns, and quills. Although the classification is based on tactile sensation, the two varieties of keratin contain different types of proteins and are produced by different modes of biosynthesis. Typically, soft keratin contains less sulfur than does hard keratin.[1]

Keratin and keratin-containing tissues have also been classified on the basis of their high-angle X-ray diffraction pattern. Soft keratin generally results in a poorly developed pattern. Certain hard keratins give a pattern consisting of two broad halos with maximum spacing of 4.5 Å and 9.5 Å. The terms α-keratin, β-keratin, feather keratin, and amorphous keratin are used to denote classification on the basis of the X-ray diffraction pattern.

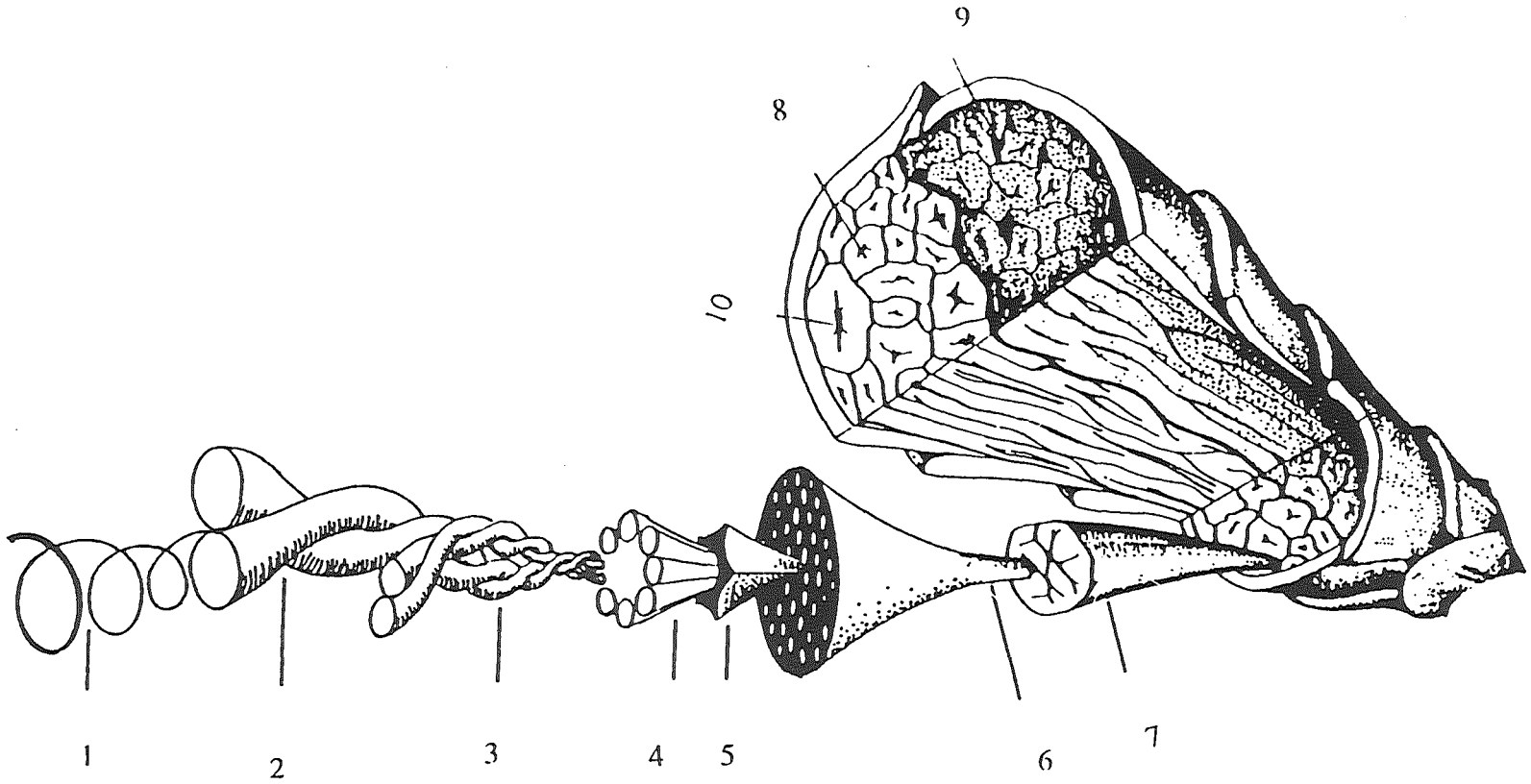


Figure 1.1. Morphology of wool fibre

1. α-helix, 2. Coiled coil dimer, 3. Tetramer of coiled coil, 4. Intermediate filament (IF) protein, 5. Intermediate filament associated protein (IFAP), 6. Macrofibril, 7. Cortical cell, 8. Paracortex cell, 9. Orthocortex cell, 10. Nuclear remnants.

Wool is a member of the class of hard α-keratin. Morphology of wool fibres has shown that fine wool fibres contain two types of cells, flattened, external cuticle cells and long, polyhedral cortical cells (Figure 1.1 Morphology of wool fibre[2]). The cuticle cells consist of three layers - epicuticle,[3] exocuticle, and endocuticle - and overlap in the longitudinal direction of the fibre rather like tiles on a roof. They are separated from one another and the underlying cortex by a cell membrane complex similar to that which separates the cortical cells from one another. The cortex is divided into two sections called the orthocortex and paracortex. The structure within each cortical cell is complex, since apart from the remains of the cellular apparatus of the once-living cell labelled 'nuclear remnants' (Figure 1.1), there are successively smaller structures, called macrofibrils, interfilament material (also called matrix), and intermediate filament (also called submicrofibrils or protofibrils). The diameter of wool fibre is generally in a range from 10 to 80 μm.

The cortical cell constitutes by far the largest amount of the fibre (about 86.5% in fine wool fibre) and is responsible for many of its important physical properties, such as elasticity. The length and width of the cortical cell is typically 95 μ and 5.5 μ respectively. It appears that the cortical cells are many-sided polyhedra which pack together without leaving any free space. The orthocortex and paracortex are approximately hemi-cylinders wound around each other helical is in phase with the crimp of the fibre. There are many differences between the orthocortex and paracortex. One of the differences is that the microfibril structure is different in the two cortices. The arrangement of the microfibril structure is much more regular in the paracortex than in the orthocortex.[4]

The macrofibrils represent aggregates of microfibrils. A single macrofibril is about 15 μ long and 0.3 μ in diameter. The scanning electron micrographs of macrofibrils have proved that the structure of macrofibrils is similar in shape to cortical cells, with finger like projections at their ends and occasionally along their length. Some of the macrofibrils have a twisted appearance.[5]

A molecular model of wool was first suggested by Astbury fifty years ago.[6] That model was a two-phase structure for microfibrils of keratin fibres, consisting of crystallites which give rise to the specific X-ray diffraction pattern (α-helical pattern of intermediate filament), embedded in an amorphous matrix of high sulfur content. The latter are now termed intermediate filament associated proteins (IFAP). Covalent disulfide bonds weld the constituent proteins into a mechanically stable structure of wool microfibrils.[7] X-ray diffraction shows the basic structure of microfibrils is constant and common for all α-keratin. Electron microscopy of stained cross sections of microfibril confirmed the two-phase structure of microfibrils.[8] The diameter of microfibrils was determined from electron microscopy to be 73 ± 1 Å. The packing of microfibrils in the paracortex of fine wool can be of various geometric types, but in some areas approximates to centred hexagonal packing.[9] Corresponding values for the centre-to-centre distance between microfibrils of the hexagon are 88 ± 1 Å. The appearance of layers or sheets of microfibrils, particularly in the orthocortex can be due to the tilting of microfibrils.[10]

The submicrofibril protein of hard α-keratin of wool fibres belong to the class of intermediate filament (IF) proteins[11] and are considered to be a special set of epithelial cytokeratins.[12,13] The question of the nature of the intermediate filament and its organisation within the microfibrils is probably the most fascinating and elusive problem in the area of α-keratin structure. The interest results from the early X-ray work which showed evidence that the structural order lies mainly within the microfibril and gave rise to the postulation of the coiled-coil α-helix by Crick[14] and Pauling.[15] This basic structure of intermediate filament proteins is now well-proved.

Since the physical properties of wool fibres are largely determined by the microfibrils and intermediate filament structures, the structural analysis of microfibrils and intermediate filaments of keratin is of economic and practical interest.

The intermediate filament of the hard α-keratin has a well-oriented protein structure and this feature allows high quality X-ray diffraction patterns to be obtained. However, the extent of disulfide bonding has greatly limited chemical investigation and

few amino acid sequences exist to complement the X-ray data.[16] Therefore, the problems posed by wool have proved to be extraordinarily complex.

Intermediate filament proteins exist not only in α-keratin, but also in non-keratin tissues. About 40 different intermediate filament proteins have been described so far.[17] Within this group of proteins five families can be recognised by immunological and nucleic acid hybridisation techniques.[18,17] Many of the recent advances in keratin structure have arisen from investigations into non-keratin members of the IF family of proteins. These non-keratin proteins have yielded a wealth of complementary data and have permitted new insights into the structure of keratin proteins.

1.1 Proteins in wool

The hard α-keratins of wool fibres comprise intermediate filament structures embedded in non-fibrillar matrix (IFAP) as described above. The task of preparing and purifying soluble derivatives of keratin protein has proven to be difficult. The most widely studied type of derivative has been prepared by reduction of the disulfide linkages in disaggregating media such as 8M urea followed by alkylation with iodoacetic acid to give sulfide (S)-carboxyl-methyl derivatives.[19] Such reduced S-carboxy-methylated extracts can be readily fractionated into three families of proteins on the basis of amino acid composition, conformation, and structural origin; two of these originate in a ground substance or matrix (IFAP) of keratin while the third is derived from microfibrils. The matrix contains two distinct groups of proteins, one characterised by having a higher sulfur content than whole keratin (termed high-sulfur proteins) and the other by a high content of glycine residues (glycine-rich proteins). All the glycine-rich proteins so far studied contain cysteine and an unusually high proportion of aromatic residues, particularly tyrosine. The microfibrils are ordered aggregates of a family of proteins that have a lower sulfur content than the whole keratin (termed low-sulfur proteins).

Chromatographic studies of the microfibrillar proteins from wool fibre suggest that there are eight low-sulfur protein species, divided into two classes of four each. Based on the classification of the intermediate filaments (IF), these two classes of wool

microfibril proteins are divided into Type I containing components 8a, 8b, 8c-1 and 8c-2, and Type II containing components 5, 7a, 7b and 7c.[20] The wool microfibril proteins have molecular weights in the range 45-58 KD[21] and isoelectric points from 4.7 to 5.4.[22] There are 903 residues in the intermediate filament (component 8c-1 and 7c) of the microfibril protein of α-keratin of wool fibre in which 7c has 491 residues and 8c-1 has 412 residues. The coiled-coil rod domain contains 622 residues, the N-terminal domain contains 164 residues and C-terminal domain contains 117 residues (see Table 1.1). The number of residues in the intermediate filament associated proteins (IFAP) is unknown in wool fibre and varies markedly in different keratins.[23] Because of the similarities in the physical and chemical properties of microfibril proteins and the fact that fractionation experiments must be carried out in disaggregating media, isolation of single purified microfibril proteins is difficult. Nevertheless, four such proteins (8a, 8c-1, 7c and 5) have been obtained from the microfibrils of wool fibre, two from each class, and extensive amino acid sequence data determined for each of them.[24]

The complete amino acid sequences including both end domains of component 8c-1 (type I)[25] and 7c (type II)[26] have been published. A detailed secondary structural analysis of the component 8c-1 has also been reported.[27] Modern DNA techniques have enabled the nucleic acid sequences corresponding to keratin proteins to be determined such that a wealth of primary structure data has become available for study and this has precipitated exciting developments.

The analysis of the homology of the amino acid sequence in the type I and type II chains shows a high degree of similarity of residues exists in the rod domain segments within each of the keratin chain types.[28] Component 8a (type I) lacks the C-terminal 30 amino acid residues of 8c-1 but the sequence is otherwise highly homologous with that of 8c-1 (Table 1.1). At least 92% identity exists between the sequence of the rod domain segments of two Type I components (8a and 8c-1). Comparison of the two type II sequences 5 and 7c shows that at the N-termini and C-termini there is essentially no homology. At least 90% identity exists between the sequences of the rod domain segments of components 5 and 7c (Table 1.1).[29]

Crewther[29] also compared a Type I 8c-1 and a Type II 7c of hard keratin protein and noted about 30% identity. There is therefore essentially little homology between the Type I and the Type II protein. Amino acid sequence similarity among members of each keratin sub-family varies. Generally the α-helical rod domains of IF proteins are 50-70% identical and the head and tail regions are less similar.[30] The coiled-coil rod domain has high overall conservation of amino acid residues. This probably is typical of intermediate filament protein from a number of sources.[31]

Table 1.1. Comparison of wool IF protein sequences[32]

	N-term.	1A	L1	1B	L12	2A	L2	2B	C-term.
Comp.7c	109	35	10	101	17	19	8	121	71
Comp.5	122	35	10	101	17	19	8	121	44+
Identities	83	31	8	90	15	15	8	113	21
Comp.8c-1	55	35	11	101	16	19	8	121	46
Comp.8a	55	35	11	101	16	19	8	121	16
Identities	48	35	10	86	16	19	8	114	8

Sequences are compared within classes. Numbers in each column are of residues in that segment. Note that the sequence of component 5 C-terminal is incomplete.

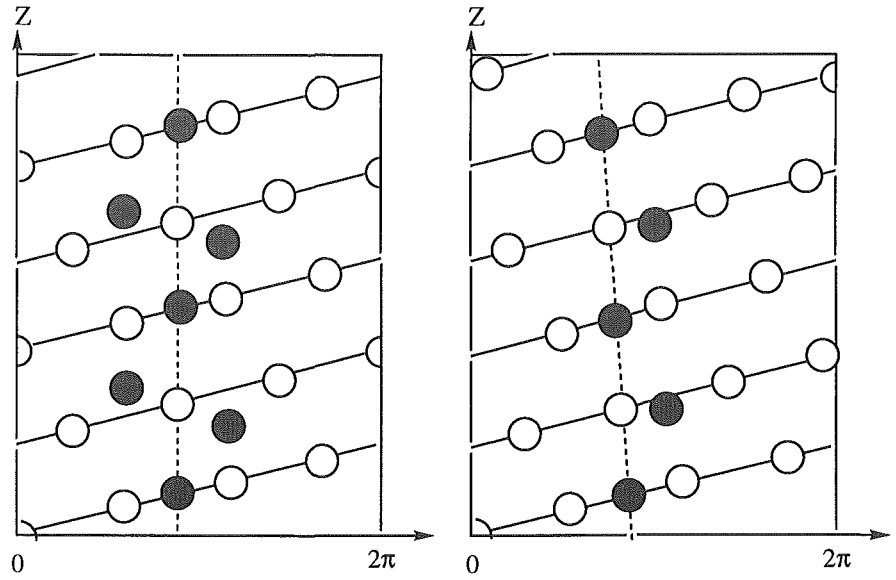
1.2 The coiled-coil model

The α-helix model of proteins proposed by Pauling and Corey[33] satisfactorily accounts for a meridional reflection of spacing about 5.4 Å in the X-ray diffraction pattern of synthetic polypeptides in the α form. However, this simple model does not predict a meridional reflection of spacing about 5.15 Å as observed in the α-pattern[34] obtained from the so-called keratin-myosin-epidermis-fibrinogen (k-m-e-f) group of fibrous proteins. Crick,[14] and Pauling and Corey[35] independently suggested that this

observed reflection could be accounted for if the axes of the α-helices in these materials were distorted so as to follow a long-pitch helix. The resulting polypeptide chain conformation was termed a coiled-coil.

The distortion was envisaged by Pauling and Corey[35] to result from a repeating sequence in which the individual residues formed hydrogen bonds of slightly different lengths. The origin of the distortion was considered to be intrahelical. Crick[14] attributed the distortion to interhelical interactions.

The radial projection of an α-helix with a unit twist t = 4π/7 has exactly 3.5 residues per turn (see Figure 1.2). If the side-chains are considered as "knobs" or "holes", the pattern of knobs and holes on the surface of the helices would allow two such helices to be brought together, with their axes parallel, in such a manner that the knobs on one helix mesh with the holes in the other.



(a) (b)

Figure 1.2. Radial projections of α-helices with (a) 3.5 residues per turn and (b) 3.6 residues per turn. The dark circles show the pattern of contacts formed by a second α-helix when brought into contact with the first. The rotation and the translation of the helices has been chosen for optimum "knob-hole" packing. In (b) "knob-hole" packing can be attained over a short length if the axis of the second helix is inclined to the axis of the first helix.

Crick[14] pointed out that the same type of knob-hole packing could be achieved with α-helices having other values of t if the axes of the helices were mutually inclined at an appropriate angle and coiled around each other. The greater the departure of t from the value of $4\pi/7$, the greater is the tilt required to maintain knob-hole packing.

Fraser[1] developed a simple relationship connecting Δt, where $\Delta t = t - (4\pi/7)$, the unit height h, the radius of the coiled-coil r_0, and its pitch P:

$$P = (2\pi/\Delta t)[h^2 - (r_0\Delta t)^2]^{1/2} \qquad 1.1$$

In all known cases t is less than $4\pi/7$ and so Δt and P are negative. The equation (1.1) is formulated by Fraser for right handed α-helices. A negative value for P corresponds to a left-handed major helix.

The coiled-coil structures of fibrous protein are also found in globular and membrane proteins.[36] For example, Landschulz[37] proposed that the leucine zipper found in some DNA-binding transcription activator proteins was α-helical and dimeric in structure. The structural data from the X-ray[38,39] and NMR[40] studies showed leucine zipper is a coiled-coil homodimer with 33 residues in each chain. A membrane protein colicin A has been shown[41] to consist of a bundle of ten α-helices in a coiled-coil. More than 200 proteins that probably have coiled-coil structures are known, including α- and β- tubulins, flagelins, G protein b sub-units, some bacterial transfer RNA synthetases, and members of the heat shock protein family.

1.2.1 The coiled-coil quations

The equations of the coiled-coil were described by Crick.[14] The coiled-coil helix (major helix) is expressed as;

$$x = r_0\cos(\omega_0 t + \varphi_0)$$
$$y = r_0\sin(\omega_0 t + \varphi_0) \qquad 1.2$$

$z = P\omega_0 t/2\pi + z_0$

and the minor helix can be expressed as;

$$x' = r_1\cos(\omega_1 t + \varphi_1)$$
$$y' = r_1\sin(\omega_1 t + \varphi_1) \qquad 1.3$$
$$z' = 0$$

where x, y, z are the coordinates of the coiled-coil helix (major helix) and x' y' z' those of the minor helix. The α-helix has a radius r_0 and a repeat distance of P in the z direction. The pitch angle of the major helix, α, is given by $\tan\alpha = 2\pi r_0/P$.

A mathematical transformation of Cartesian system in 3D from the minor helix into the major helix can produce a further expression for the coordinates of atoms in the minor helix in terms of the major coordinates frame as;

$$x = r_0\cos(\omega_0 t + \varphi_0) + r_1\cos(\omega_1 t + \varphi_1)\cos(\omega_0 t + \varphi_0) + r_1\cos\alpha\sin(\omega_1 t + \varphi_1)\sin(\omega_0 t + \varphi_0)$$
$$y = - r_0\sin(\omega_0 t + \varphi_0) - r_1\cos(\omega_1 t + \varphi_1)\sin(\omega_0 t + \varphi_0) + r_1\cos\alpha\sin(\omega_1 t + \varphi_1)\cos(\omega_0 t + \varphi_0)$$
$$z = P(\omega_0 t/2\pi) + z_0 + r_1\sin\alpha\sin(\omega_1 t + \varphi_1) \qquad 1.4$$

The parameter φ_1 can be calculated from

$$\varphi_1 = \varphi_s + \{(N_1 - N_0)/M\}\varphi_M \text{ with } \varphi_M = 2\pi M z_s\cos\alpha/P \qquad 1.5$$

Where ;

r_0 = the radius of the major helix,

r_1 = the radius of the minor helix,

φ_0 = the phase angles of the axis of the minor helix,

φ_s = the phase angle of atom on the minor helix,

ω_0 = the rotation angle of atom in major helix (°),

ω_1 = the rotation angle of atom in minor helix,

M = the number of residues in a repeat distance P (pitch),

t = 1, 2, 3,.....M,

α = the pitch angle,

p = the pitch or the repeat distance on z axis direction in major helix,

z_s = the starting height of atoms on minor helices,

N_1 = the number of turns of the minor helix,

N_o = the number of turns of the major helix.

The structure of the coiled-coil helix repeats after a distance c ($c = N_oP$) in the Z direction. The parameters used in the equations of the coiled-coil helices were also proposed by Crick based on the X-ray pattern of hard α-keratin as;

Turns of the major helix,	$N_o = 1$.
Turns of the minor helix,	$N_1 = 36$.
Atoms in the set,	M = 126.
Repeat distance,	C = 186 Å
Radius of major helix,	r_o = 5.2 Å.

1.2.2. The Fourier transform

The Fourier transform of helical structures has certain characteristic features and an understanding of these features is useful for the interpretation of diffraction patterns obtained from fibrous proteins. The theory of diffraction by helical structures was first given by Cochran et al[42] 45 years ago and has been considerably amplified by Klug.[43]

Crick[35,37] described a method for calculating the Fourier transform of a coiled-coil of the fibrous proteins and alternative derivations have been discussed by Lang[44] Subsequently Pardon[45] showed that certain of the assumptions made in these treatments were not strictly correct but it was considered that the error involved in using Crick's method would be small since the distortion of the α-helix was relatively slight.

Fraser[46] presented a method of calculating the Fourier transform, which is not subject to these uncertainties, and depends on the fact that coiled-coils contain a repeating unit of seven residues which is arranged with helical symmetry. Exact calculations can therefore be carried out using the following formula for a single set of points of a coiled-coil helix.[48]

$$C(R, \psi, l/c) = \sum_p \sum_q \sum_d \sum_s J_p(2\pi R r_o) \times J_q(2\pi R \bar{r}_1) J_s(2\pi(l/c) r_1 \sin\alpha) \times J_d(2\pi R \Delta) \exp\{i[p(\psi - \varphi_o + \pi/2) + q(-\psi + \varphi_o + \varphi_1 + \pi/2) + s(\pi + \varphi_1) + d(\psi + \varphi_1 - \varphi_o + \pi/2) - m'\varphi_M + (2\pi z_o l/c)]\} \quad 1.6$$

subject to the restriction that,

$$N_o p + (N_1 - N_o)q + N_1 s + (N_1 + N_o)d = l + Mm' \quad 1.7$$

where p, q, s, d, and m' may take any integral value, positive or negative. R, ψ, l/c (=Z) are cylindrical coordinates of the scattering transform in reciprocal space; r_o, φ_o, z_o are the polar coordinates of origin of the rotating frame at t = 0; c is repeat distance and l is an integral number; $\bar{r}_1 = r_1(1+\cos\alpha)/2$ and $\Delta = r_1(1-\cos\alpha)/2$. r_1, φ_1, φ_M, α, N_1, N_o and M are defined in the equation 1.5. Equation 1.6 may be simplified since $\alpha \approx 10^\circ$ and so Δ is small and d is restricted to zero, therefore the Bessel function $J_d(2\pi R\Delta) = J_0(0) = 1$.

The Fourier transform C(R, ψ, Z) is non-zero only on the layer-line (l is the number of the layer-line) since the structure is periodic in the z direction but non-periodic in the other two directions. In the case of an infinite coiled-coil with perfect helical symmetry the diffraction pattern would be confined to a set of layer lines with Z coordinates.

The nature of the solution of equation 1.6 is simply described in term of parameters m and λ which are related to l by

$$Z(m, l) = l/c = m/h + \lambda/P. \qquad 1.8$$

where P is the pitch of the coiled-coil, h is the height of the seven residues asymmetric unit, and m and λ are integers. The equation 1.8 emphasises the fact that the coiled-coil can be regarded as a simple helix of pitch P with asymmetric units of seven residues distributed at vertical intervals of h. The Fourier transform will consist of branches emanating from the origin, and a series of points on the meridian with a distance of m/h from the origin, where $m = \pm 1, \pm 2$, etc. The separation between each layer line and the next in a branch will be 1/P.

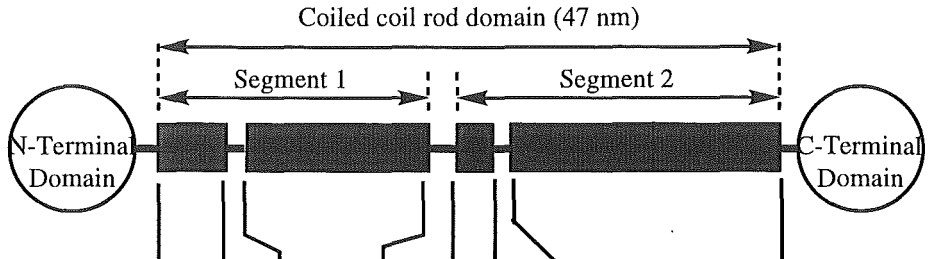
From the Fourier transform, Crick explained that the 5.15 Å meridional reflection would be described by $m = 2$, $\lambda = 0$, the 1.5 Å meridional reflection by $m = 7$, $\lambda = 0$ and the 10 Å equatorial reflection by $m = 0$, $\lambda = 0 \pm 1, \pm 2$ etc. Thus the general features of the observed X-ray pattern of the coiled-coil proteins can be predicted and explained by the Fourier transform.

1.3 The models of intermediate filament protein

1.3.1 Two chain model

Many features of keratin IF structure have now gained wide acceptance. These include the observation that the chains in both the hard and soft keratin form a central rod domain. An amino acid sequence assumes the central rod domain contains a succession of seven-residue peptides (heptads) of the kind of $(a\text{-}b\text{-}c\text{-}d\text{-}e\text{-}f\text{-}g)_n$ with a high probability of adopting an α-helical conformation, enclosed by non-α-helical N- terminal and C-terminal domains.[47,48,49,50] The heptad substructure and α-helical rich features of the rod domain are indicative of a coiled-coil conformation characteristic of all α-fibrous proteins. The rod domain is about 47 nm long and consists of two segments (segment 1 and 2), each about 22 nm in length. Segment 1 (1A-L1-1B) comprises two heptad-containing segments 1A and 1B separated by a variable length. The number of residues in the variable length region is different for different fibrous proteins and is referred to as the non-coiled-coil link L1 region. Likewise segment 2 (2A-L2-2B) contains a pair of

heptad-containing segments 2A and 2B which in this case are separated by a constant length (eight residues) non-helical link L2. Segment 1 and 2, in turn, are connected by a non-helical link L12 (See Figure 1.3)



	End	1A	L1	1B	L12	2A	L2	2B	End
8c-1	55	35	11	101	16	19	8	121	71
7c	109	35	10	101	17	19	8	121	46

Figure 1.3. The model of the coiled-coil rod domain in the fibrous proteins

Strong evidence has been presented that keratin molecules consist of two heteropolymers,[51,52] and that the two chains in the coiled-coil are both parallel and in axial register.[53,54,55,56] It is clear that the rod domain of wool protein is made up from the intermediate filament proteins of Type I and Type II chains such as chain 8c-1 and 7c.

From various analyses of the secondary structure of IF proteins, it is evident that all IF chains possess a central α-helical rod domain flanked by end domains of widely different size and chemical character. Thus, differences in the size and properties of IF proteins are due almost entirely to the variability of their end domains. The end domains have commonly high glycine content configured in tandem quasi peptide repeats of the form aliphatic - $(glycine/serine)_n$.[57]

1.3.2 Four chain model

For wool the evidence is convincing that there is a four-chain complex consisting of a pair of two-chain coiled-coil molecules.[13,58] A four-chain structural unit has been isolated from hard and soft keratin.[59] The pair of these two coiled-coil molecules is considered the smallest, stable subfilamentous oligomer that can exist in solution.[60] The

models of keratin intermediate filament protein invoke the following hierarchical steps in assembly from constituent protein chains:[61,62,63] (a) two compatible chains align themselves in parallel and in axial register to form a two chain coiled-coil molecule;[64] (b) a pair of these two coiled-coil molecules form a four chain complex;[65] and (c) a number of these four chain complexes ranging from about 7 to 8 in a cross-section of microfibril,[66] associate largely by intermolecular ionic interactions of the rod domains to form the intact keratin IF protein.[67,68]

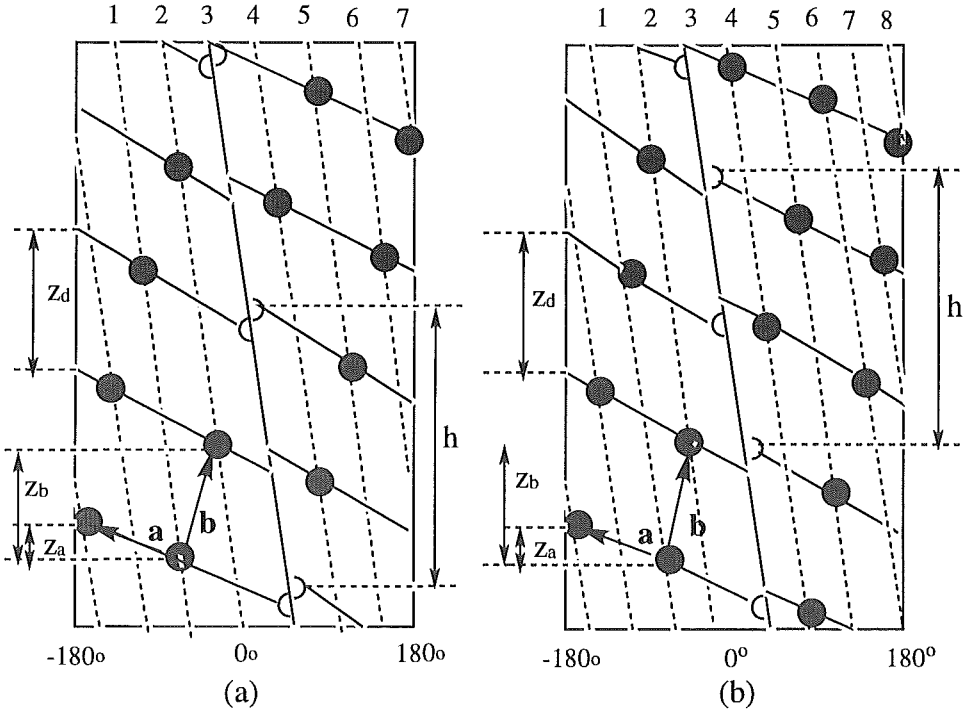


Figure 1.4. (a) Radial projection (not to scale) of the α-keratin IF surface lattice based on seven repeating units per 47 nm (h). Lattice vectors **a** and **b** have axial projections of z_a (7.42 nm) and z_b (19.79 nm), respectively. When wrapped around a cylindrical surface the lattice does not form continuous helices, and the dislocation is indicated by a full line. The filament may be regarded as being made up of seven subfilaments (shown black dotted), as indicated by the number at the top of the diagram. There is a stagger of z_b between adjacent subfilaments. The filament has true helical symmetry[69] with a basic helix of pitch -344.7 nm, a unit height of 47 nm, and a unit twist

-49.1° (b) As in (a) except that lattice is based on eight repeating unit per 47 nm. Excellent agreement with linear mass estimates[70,71] is obtained. The agreement in (a) is less satisfactory and the possibility of a central subfilament has been proposed by Fraser.[70]

Fraser and colleagues[72] have documented an extensive series of meridional reflections in the diffraction pattern from the highly oriented keratin of the porcupine quill. They indexed these data in terms of a helical surface lattice with an axial repeat of 47 nm containing seven or eight steps on a basic helix of 22 nm pitch. They proposed that each of these steps or lattice points is occupied by a four-chain unit. On the basis of the estimates[73] of the linear mass distribution in IF protein using scanning transmission electron microscopy (STEM), it is believed that a single four-chain unit is associated with each surface lattice point of the microfilament of keratin[74] (see Figure 1.4).

The way in which these two-stranded coiled-coil molecules aggregate to form higher-order polymers of four chains in IF is, however not certain. Crewther[64] proposed that the number of possible ionic interactions between neighbouring segments would be a major determinant in their alignment. He pointed out that the maximum numbers of favourable ionic interactions occur when the segments 1B and 2B of neighbouring molecules are adjacent. Thus, there are four simple possibilities: neighbouring molecules could be parallel or anti-parallel, in exact axial register, or half-staggered. From chemical, electron microscope and theoretical analysis it has been shown that each four-chain complex consists of a pair of anti-parallel molecules rather than of a pair of parallel molecules.[73] Therefore, the four possibilities are reduced into two. These two possible modes of association of a pair of anti-parallel molecules have been suggested. In the first the molecules are almost completely overlapped[74] and in the second the molecules are approximately half-staggered.[75]

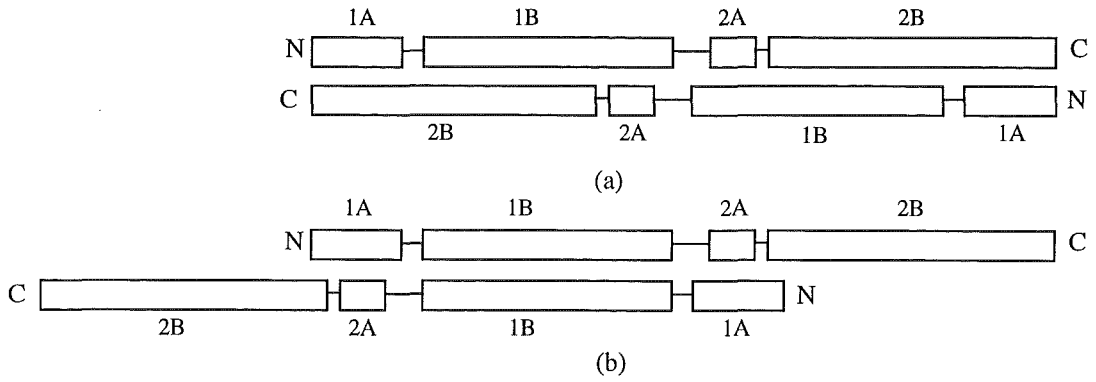


Figure 1.5. The model of assembly of coiled-coil rod domains in the tetramer of IF protein (the end domains are not drawn). (a) The two coiled-coil rod domains in the tetramer of IF are anti-parallel and in axial register. (b) The two coiled-coil rod domains are anti-parallel and staggered.

Both models may be possible in native IF proteins. Based on a theoretical analysis and calculation of the surface lattice of IF proteins of hard α-keratin, Fraser et al[76] suggested that the tetramer model had a high potential for establishing favourable ionic interactions and disulfide bonds between two coiled-coil ropes[77] in an anti-parallel and staggered conformation. Potschka et al.[78] confirmed the two coiled-coil rod domains in the tetramer are staggered by approximately 15 nm in desmin protofilaments by using high-resolution gel permeation chromatography (GPC) and electron microscopy techniques.

The disulfide bond is also an important factor to be considered in the assembly of tetramer. A theoretical study[80] of potential disulfide bonds between the rod domains of wool microfibrillar proteins showed that the only possible disulfide bonds were between the 2B segments of different tetramers. The tetramers are pairs of heterodimers formed by parallel in register association of the monomeric IF protein. This study identified potential disulfide bonds between the 2B segments of components 7c and 8c-1. It was suggested that three out of the four cysteines in the 2B segment of each of these proteins were involved giving a total of six inter-rope disulfide bonds. In order to form these six disulfide bonds a stagger of thirteen residues must be postulated between 2B segments of

adjacent tetramers. However, when the sequences of components 8a (Type I) and 5 (Type II) are examined in the same way it is found that for component 8a only one of the four cysteines in the 2B segment of 8c-1 is present while for component 5 three out of the four cysteines in the 2B segment of 7c is present. Thus for the combination of components 8a and 5 it is possible that only two disulfide bonds are formed in this region. It may be however, that in the native filaments, components 8a and 5 are not associated. The concept suggested by these considerations is that of specificity of association determined by the potential for disulfide bond formation. This cannot be explored further until sequence data is available for the other wool microfibrillar proteins.

Outside the helical segments, disulfide bonds cannot be easily predicted as the protein conformation in these segments is not understood. However, it should be pointed out that the positions of cysteines throughout the entire homologous regions of the two pairs of proteins are not highly conserved. This suggests that the proteins in these pairs are not functionally equivalent. Because the conformation of the linking segments of the coiled-coil rod domain is not known it is still not possible to suggest a detailed structure for the tetramer repeat of wool IF protein. [79]

1.4 Heptad repeat of intermediate filament protein

Crick[14] pointed out that a key 5.15 Å meridional reflection in the X-ray pattern could arise from layers of side-chains between interlocking α-helices. He showed that α-helices with about 3.5 residues per turn could mesh together locally if their axes were inclined to each other at about 18°. Moreover, if the amino acid sequences have a seven-residue 'heptad' repeat, in which *a* and *d* are generally apolar residues (Figure 1.6), then an apolar surface stripe would arise inclined around the axis of each α-helix. The physical basis of long range super-coiling, then, is a regular pattern of 'knobs-into-holes' packing of apolar side-chains, which acts to stabilise the structure when two or more helices interact.

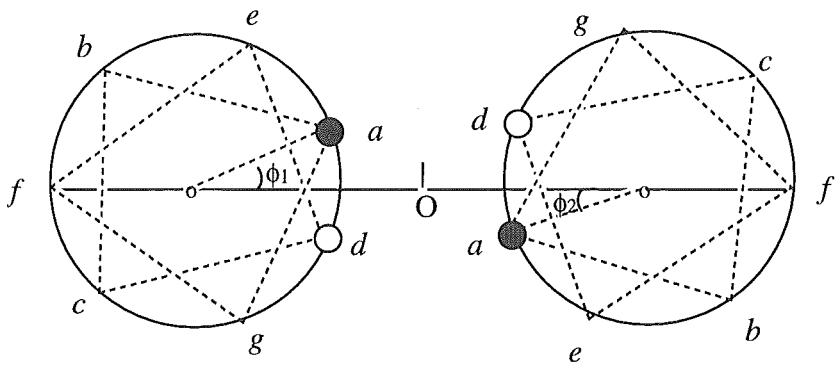


Figure 1.6. Seven positions in heptad repeats of residues in coiled-coils

The first α-fibrous protein sequence to be determined was in fact that of tropomyosin. The two chains of 284 residues in each chain in this protein were shown to have an unbroken periodicity, 40 heptads in length.[80] Other representative proteins in the k-m-e-f class have been sequenced and also found to contain long runs of heptad repeats in these IF proteins.[81,82]

The coiled-coil is a structure stabilised by the interactions between α-helices. This coiling not only optimises interchain packing of apolar residues, but also allows specific interchain ionic interactions to be made which define the relative chain alignment and direction. Thus the essential aspect of the coiled-coil structure is not the bending of the axis of the α-helix in the supercoiled conformation, but the systematic side-chain interactions that are permitted. It is believed that the heptads of the residues in the two helical chains are one of the essential properties of the coiled-coils. Most investigators have shown that the *a* and *d* positions in heptads in the coiled-coil helices are mostly occupied by apolar residues,[83] leucines are especially located at the *d* position.[84] In the fibrous proteins, more than 34% of leucine residues are located in the *d* position.[85] The heptads of residues are found not only in IF proteins but also in membrane proteins and in globular proteins.[86] For example, in the protein 2ZTA, the percentage of leucine residues occupying in the *d* position is as high as 67% (8/12) in both helical chains respectively.

Except gly and pro the apolar residues (Ala, Ile, Val, Leu, Phe, Pro, Met and Gly) are good helix-formers in protein.[87,88] However, it is not clear why the Leu residues have a particularly preference on the *d* position among these apolar amino acid residues. Cohen and Parry[89] explained that the β-branch of the side-chain can have an influence on the packing of the *a* and *d* positions. A side-chain in the *d* position points directly into the interface, whereas a side-chain in the *a* position points out from the interface. This fact explains the observed preference in the fibrous proteins for β-branched side-chains (Ile or Val) in the *a* position where they can redirect part of their side-chain back into the core. It also explains the tolerance in all coiled-coils for the Lys residues in the *a* but not the *d* positions.

The presence of branching in the amino acid residues of position *a* and *d* determines the state of oligomerization of the α-helical chains.[90] Ile in the *a* and Leu in *d* yields the zipper-like two stranded structure. Ile in both *a* and *d* position yields a three-stranded coiled-coil structure; and Leu in *a* and Ile in *d* give rise to a four-stranded assembly. Thus β-branched side-chain of residues tend to destabilize particular modes of packing. These results are consistent with the sequences of known two- and three-stranded α-fibrous proteins.[91] However, this point is not universally accepted and further data will be required to settle this point.

Since the *a* and *d* positions are on the inter-face of two coiled-coil helical chains (Figure 1.6) or in the "core" positions in the coiled-coil structures, the non-polar residues are expected to be located in the *a* and *d* positions because the residues of non-polar residues mostly prefer knob-hole interactions and are hydrophobic. Charged amino acid residues prefer the outside of the helices such as *b, c, f* positions and are directed towards the polar solvent.

In the models of the coiled-coils, the heptad positions of the residues of two or more α-helical chains can be defined by independently rotating two phase angles (ϕ_1 and ϕ_2) of the minor helix as shown in figure 1.6. If the phase angles (ϕ_1 and ϕ_2) of the residues are rotated with respect to each other, the residue positions in the heptad repeats

can be defined by meeting the criteria according to which most Leu residues are located at the *a* or *d* positions in the two chains of the coiled-coil structure.

1.5. The strategy of modelling studies

Computer modelling of protein structures has been growing rapidly with the development of computer techniques and technology in recent years. Computer modelling techniques can calculate 3D structures of a protein from the experimental data of X-ray and NMR methods.[92] The proteins for which a 3D structure can not be determined by X-ray or NMR can be simulated and predicted from the presented information of the protein. The properties of protein structures such as interaction energy, hydrogen bonding, disulfide bonds etc. can be predicted by computer based studies.

Several investigators have made progress on modelling coiled-coil proteins since the structure was first proposed. Fraser[48] was the first to calculate the coordinates of the thirty-five atoms in a coiled-coil model. He used a radius 5.2 Å and a pitch 186 Å. He examined the distortion of individual bond lengths and inter bond angles in the coiled-coil structure. He pointed out that in a coiled-coil the distortion would not be uniformly distributed. It is likely that most of the distortion will be accommodated by changes in the torsion angles of the backbone and in the interbond angle rather than by changes in bond lengths. His model was built in the manner of a right handed coiled-coil with left handed α-helices because at that stage, it was not clear that the α-helix of L-amino acids was right handed and the supercoil left handed.

Parry and Suzuki[93] established a coiled-coil model of poly-L-alanine with a pitch of 186 Å and radius of 5.2 Å by using Crick's equations and corresponding parameters. The model was for a left handed supercoil and for right handed α-helices. The energy associated with the coiled-coil model was estimated by using simple non-bonding interactions involved in Van der Waal and electrostatic interactions. The potential energies in straight and deformed α-helices were compared. The energy of deformation from the straight α-helix was small, as Crick[14] had estimated.

McLachlan[94] constructed a space-filling model of the coiled-coil structure of Murein Lipoprotein, a membrane of homodimer protein with 58 residues in each chain.[95] The pitch was set at 186 Å and the structures confirmed that the coiled-coil model is stereo-chemically reasonable. Energy calculations for a series of coiled-coils with different radii suggested that the best structure is one with the helix axes 8.25 Å apart i.e. a radius of the coiled-coil of 4.125 Å.

Nilges[96] developed an approach for the modelling of coiled-coil structures by molecular dynamics and applied the method to the protein CAP,[97] part of which forms a coiled-coil, and to the leucine zipper GCN4.[98] Both models were built with an infinite pitch (the two helical chains are almost parallel) and a radius of 5.2 Å. The average r.m.s. deviation of the model structure of CAP, compared with the X-ray structure is 0.7 Å for the backbone atoms if the first and last three residues of backbone are ignored and 1.3 Å for the buried side-chain atoms and 2.7 Å for the outside side-chain atoms. The r.m.s. deviation was not obtained for leucine zipper because the X-ray crystal structure of 2ZTA was not available at that time. The radius of 5.2 Å of coiled-coil in the model of leucine zipper reduced to 4.9 Å in the final model.

In the past few years structural information for proteins has been growing in an exponential-like fashion. Determinations of protein structures by X-ray crystallography and NMR spectroscopy have contributed significantly to this expanding database of knowledge. Crystallographic information is complied in the Brookhaven Protein Databank (PDB)[99] and for coiled-coil protein crystals, the data entries are limited to a few coiled-coils such as leucine zipper (2ZTA)[100], 1ROP,[101] and tropomyosin.[102]

Based on the structural information, several investigators have developed computer-aided model building algorithms for the purpose of constructing all atom structures. The methods can be divided into generation of all-atom structure from the segment matching of homologous proteins whose structures are in the PDB and the generation of the all atom structures of proteins from C^{α} coordinates. The accuracy of the methods can be established by comparison of the model structures with the corresponding X-ray structures.

Our strategy for the modelling of the coiled-coil structure in wool protein is discussed in the following chapters. The generation of the C^{α} coordinates for the coiled-coil model is carried out by using Crick's equations. The generation of all atomic model is effected by Monte Carlo Protein Building (MCPB) (see Chapter 2 and 3) with simulated annealing techniques (MCPBA) (see Chapter 4). This building process can involve varied initial parameters of the coiled-coils. The method was tested on a set of nine proteins and a small coiled-coil protein GCN4 and compared the model structures with the corresponding X-ray crystal structures (Chapter 4). From the model, a series of properties of the coiled-coil rod domain of wool protein is examined (Chapter 5); The interaction between two coiled-coil helical chains, the residues in the heptad positions along the chains, the hydrogen bonds and disulfide bonds in the side-chains, the surface and volume of the rod domains etc. are examined. The results of these studies will be given in the following chapters.

Reference

1 Fraser R. D. B., MacRae I. P. Conformation in Fibrous Proteins Academic (1973) press New York and London.

2 Maclaren J. A., Milligan B. Wool Science: The chemical reactivety of the wool fibre, Science Press (1981) New South Wales Australia. Zahn H. (1988) *Chemica* **42,** 289-297.

3 The existance of an epicuticle layer is currently being debated with some scientists believing this to be part of A-layer of the endocuticle. (personal communication Dr Gill Worth)

4 Dobb M. G. *J. Text. Inst.* (1970) **61,** 232.

5 Bradbury J. H., Peters D. E. *Text. Res. J.* (1972) **42,** 471.

6 Astbury W. T., Dickinson S. *Proc. Roy. Soc. Ser. B.* (1940) **129,** 307.

7 Fraser R. D. B., MacRae T. P. *Nature* (1971) **233,** 138-140.

8 Crewther W. G., Fraser R. D. B., Lennox F. G., Lindley H. *Adv. Protein Chem.* (1965) **20,** 191.

9 Fraser R. D. B., MacRae T. P., Rogers G. E. *J. Text. Inst.* (1960) **51,** T497.

10 Johnson D. J., Sikorski J. *Int. Wool Text. Res. Conf., Proc. 3rd, Paris* (1965) **1,** 147.

11 Steinert P. M., Steven A. C., Roop D. R. *Cell* (1985) **42,** 411-419.

12 Steinert P. M. *J. Mol. Biol.* (1978) **123,** 49-70.

13 Geisler N., Weber K. *EMBO J.* (1982) **1,** 1649 -1656.

14 Crick F. H. C. *Acta Crystallogr.* (1953) **6,** 685-689.

15 Pauling L., Corey R. B. *Nature (London)* (1953) **171,** 59.

16 Rogers G. E. The Biology of Wool and Hair Chapman & Hall, London & New York (1988) p128.

17 Blumenberg M. *Mol. Biol. Evol.* (1989) **6,** 53-65.

18 Steinert P. M., Parry D. A. D. *Ann. Rev. Cell Biol.* (1985) **1,** 41-65.

19 Mies H. H., Zahn H. *J. Chromatogr.* (1987) **405,** 365-370.

20 Crewther W. C., Dowling L. M., Gough K. H., Marshall R. C., Sparrow L. G. In Fibrous Proteins: Scientific, Industrial and Medical Aspects (1980) vol.2 p 151-159, Academic Press, London.

21 Woods E. F. *Aust. J. Biol. Sci.* (1979) **32,** 423-435.

22 Marshall R. C., Blagrove R. J. *J. Chromatogr.* (1979) **172,** 351-356.

23 Fraser R. D. B., MacRae T. P., Sparrow L. G., Parry D. A. D. *Int. J. Biol. Macromol.* (1988) **10,** 106-112.

24 Rogers G. E. The Biology of Wool and Hair Chapman & Hall, London & New York (1988) p140.

25 Dowling L. M., Crewther W. G., Inglis A. S. *Biochem. J.* (1986) **236,** 695-703.

26 Sparrow L. G., Robinson C. P., McMahon D. T. W., Rubira M. R. *Biochem. J.* (1989) **261,** 1015.

27 Dowling L. M., Crewther W. G., Parry D. A. D. *Biochem. J.* (1986) **236,** 705-712.

28 Conway J. F., Parry D. A. D. *Int. J. Biol. Macromol.* (1988) **10,** 79-98.

29 Crewther W. G., Dowling L. M., Inglis A. S., Sparrow L. G., Strike P. M. Woods, E. F. *7th Int. Wool Text Res. Conf.* (1985) Tokyo I 85-94.

30 Steinert P. M., Steven A. C., Roop D. R. *Cell* (1985) **42,** 411-419.

31 Eckert R. L. *Proc. Natl. Acad. Sci. USA* (1988) **86,** 1114-1118.

32 Rogers G. E. The Biology of Wool and Hair London &New York (1988) p148.

33 Pauling L., Corey R. B. *Proc. Nat. Acad. Sci. U.S* (1951) **37,** 235.

34 Astbury W. T., Woods H. J. *Nature (london)* (1930) **126,** 913.

35 Pauling L., Corey R. B. *Nature (London)* (1953) **171,** 59.

36 Cohen C., Parry D. A. D. *Protein* (1990) **7,** 1-15.

37 Landschulz W. H., Johnson P. F., McKnight S. L. *Science* (1988) **240,** 1759.

38 O'Shea E. K., Rutkowski R., Kim P. S. *Science* (1989) **243,** 538.

39 O'Shea E. K., Rutkowski R., Stafford W. F., Kim P. S. *Science* (1989) **245,** 646.

40 Oas T. G., McIntosh L. P., O'Shea E. K., Dahlquist F. W., Kim P. S. (1990) *Biochemistry* **27,** 2892-2894.

41 Parker M. W., Pattus F., Tucker A. D., Tsernoglou D. *Nature* (1989) **337,** 93-96.

42 Cochran W., Crick F. H. C., Vand V. *Acta Crystallogr.* (1952) **5,** 581.

43 Klug A., Crick F. H. C., Wyckoff H. W. *Acta Crystallogr.* (1958) **25,** 104.

44 Lang A. R. *Acta Crystallogr.* (1956) **9,** 436.

45 Pardon J. F.*Acta Crystallogr.* (1967) **23,** 937.

46 Fraser R. D. B., MacRae T. P., Miller A. *Acta Crystallogr* (1964) **17,** 813.

47 Steinert P. M., Idler W. W., Goldman R. D. *Proc. Natl. Acad. Sci. U.S.A.* (1980) **77,** 4534-4538.

48 Steinert P. M., Parry D. A. D., Racoosin E. L., Idler W. W., Steven A. C., Trus B. L., Roop D. R. *Proc. Natl. Acad. Sci U.S.A.* (1984) **81,** 5709-5713.

49 Crewther W. G., Dowling L. M., Steiner P. M., Parry D. A. D. *Int. J. Biol. Macromol.* (1983) **5,** 267-274.

50 Parry D.A.D., Fraser R.D.B. *Int. J. Biol. Macromol.* (1985) **7,** 203-213.

51 Hatzfeld M., Franke W. W. *J. Cell Biol.* (1985) **101,** 1826-1841.

52 Eichner R., Sun T.-T., Aebi U. *J. Cell Biol.* (1986) **102,** 1767-1777.

53 Parry D. A. D., Crewther W. G., Fraser R. D. B., MacRae T. P. *J. Mol. Biol.* (1977) **113,** 449-454.

54 Woods E. F., Gruen L. C. *Aust. J. Biol. Sci.* (1981) **34,** 515-526.

55 Quinlan R. A., Franke W. W. *Proc. Natl. Acad. Sci. U.S.A.* (1982) **79,** 3452-3456.

56 Pang Y. Y. S., Robson R. M., Hartzer M. K., Stromer M. H. *J. Cell Biol.* (1983) 97:226.

57 Steinert P. M., Parry D. A. D., Racoosin E. L., Idler W. W., Johnson L. D., Roop D. R. *J. Biol. Chem.* (1985) **260,** 7142-7149.

58 Woods E. F., Inglis A. S. *Int. J. Biol. Macrmol.* (1984) **6,** 277-283.

59 Quinlan R. A., Cohlberg J. A., Schiller D. L., Hatzfeld M., Fanke W. W. *J. Mol. Biol.* (1984) **178,** 365-388.

60 Gruen L. C., Woods E. F. *Biochem. J.* (1983) **209,** 587-595.

61 Steinert P. M., Roop D. R. *Annu. Rev. Biochem.* (1988) **57,** 93-625.

62 Crewther W. G., Dowling L. M., Steinert P. M., Parry D. A. D. *Int. J. Biol. Macromol.* (1983) **5,** 267-274.

63 Steinert P. M., Steven A. C. *Nature* (1985) **316,** 767.

64 Parry D. A. D., Steven A. C., Steinert P. M. *Biochem. Biophys. Res. Commun.* (1985) **127,** 1012-1018.

65 Farnke W. W., Schiller D. L., Hatzfeld M., Winter S. *Proc. Natl. Acad. Sci. U.S.A.* (1983) **80,** 7113-7117.

66 Steven A. C., Hainfeld J. F., Trus B. L., Wall J. S., Steinert P. M. *J. Biol. Chem.* (1983) **258,** 8323-8329.

67 Fraser R. D. B., MacRae T. P. *Biosci. Rep.* (1983) **3,** 517-525.

68 Fraser R. D. B., MacRae T. P., Parry D. A. D., Suzuki E. *Proc. Natl. Acad. Sci..U.S.A.* (1986) **83,** 1179-1183.

69 Fraser R. D. B., MacRae T. P., Roger G. E. *J. Mol. Biol.* (1976) **108,**435-452.

70 Steven A. C., Wall J., Hainfeld J., Steinert P. M. *Proc. Natl. Acad. Sci. USA* (1982) **79,** 3101-3105.

71 Steven A. C., Hainfeld J. F., Wall J. S., Steinert P. M. *J. Cell Biol.* (1983) **97,** 1939-1944.

72 Fraser R. D. B., MacRae T. P. *Polymer* (1973) **14,** 61-67.

73 Geisler N., Kaufmann E., Weber K. *J. Mol. Biol.* (1985) **182,** 173-177.

74 Quinlan R. A., Cohlberg J. A., Schiller D. L., Hatzfeld M., Fanke W. W. *J. Mol. Biol.* (1984) **178,** 365-388. Geisler N., Kaufmann E., Weber K. *J. Mol. Biol.* (1985) **182,** 173-177.

75 Potschka M. *Biophys. J.* (1986) **49,** 129-130.

76 Fraser R. D. B., MacRae T. P., Suzuki E., Parry D. A. D., Trajstman A. C., Lucas I. *Int. J. Biol. Macromol.* (1985) **7,** 258-274.

77 Fraser R. D. B., MacRae T. P., Sparrow L. G., Parry D. A. D. *Int. J. Biol. Macrmol.* (1988) **10,** 107-112.

78 Potschka M., Nave R., Weber K., Geisler N. *Eur. J. Biochem.* (1990) **190,** 503-508.

79 Rogers G. E. The Biology of Wool and Hair London and New York (1988) p154.

80 Mclachlan A. D., Stewart M. *J. Mol. Biol.* (1975) **98,** 293-304.

81 Mclachlan A. D., Karn J. *J. Mol. Biol.* (1983) **164,** 605-626.

82 Doolittle R. F., Cassman K. G., Cottrell B. A., Friezner S. J., Takagi T. *Biochemistry* (1977) **16,** 11710-1715.

83 Cohen C., Reinhardt B., Parry D. A. D., Roelants G. E., Hirsch W., Kanwe B. *Nature* (1984) **311,** 169-171.

84 Steinert P. M., Parry D. A. D., Racoosin E. L., Idler W. W., Johnson L. D., Roop D. R. *J. Biol. Chem.* (1985) **260,** 7142-7149.

85 Cohen C., Parry D. A. D. *Protein* (1990) **7,** 1-15.

86 Cohen C., Parry D. A. D. *TIBS* (1986) **11,** 245-248.

87 Padmanabhan S., Baldwin R. L. *J. Mol. Biol.* (1991) **219,** 135-137.

88 Richardson J. S., Richardson D. C. *Science* (1988) **240,** 1648-1652.

89 Cohen C., Parry D. A. D. *Science* (1994) **263,** 488-489.

90 Conway J. F., Parry D. A. D. *Int. J. Biol. Macromol.* (1990) **12,** 328.

91 Conway J. F., Parry D. A. D. *Int. J. Biol. Macromol.* (1991) **13,** 14.

92 Kaptein R., Boelens R., Scheek R. M., Van Gunsteren W. F. *Biochemistry* (1988) **27,** 5389-5395.

93 Parry D. A. D., Suzuki E. *Biopolymers* (1969) **7,** 189-198. Parry D. A. D., Suzuki E. *Biopolymers* (1969) **7,** 199-206.

94 McLachan A. D. *J. Mol. Biol.* (1978) **122,** 493-506.

95 Braun V., Bosch V. *Proc. Natl. Acad. Sci. USA* (1972) **69,** 970-974.

96 Nilges M., Brunger A. T. *Protein Engineering* (1991) **4,** 649-659.

97 Weber I. T., Steitz T. A. *J. Mol. Biol.* (1987) **198,** 311-326.

98 O'Shea E. K., Rutkowski R., Stafford W. F., Kim P. S. *Science* (1989) **245,** 646.

99 Bernstein F. C., Koetzle T. F., Williams E. J. B., Meyer Jr. E. F., Kennard, O., Shimanouchi T., Tasumi M. *J. Mol. Biol.* (1977) **112,** 535.

100 O'Shea E. K., Klemm J. D., Kim P. S., Alber T. *Science* (1991) **254,** 539.

101 Banner D. W., Kokkinidis M., Tsernoglou D. *J. Mol. Biol.* (1987) **196,** 657-675.

102 Phillips G. N., Fillers J. P., Cohen C. *J.Mol. Biol.* (1986) **192,** 111-131.

Chapter Two

Assignment of secondary structure from C^{α} coordinates

Summary

A multiple regression analysis has established a non-linear relationship between the backbone dihedral angles and the C^{α} coordinates obtained from the X-ray crystal structures of fourteen proteins. The regression equations have been applied to predict specific dihedral angles of each residue in the backbone of twenty-four proteins. Overall this method (NLRDT) predicts values of ϕ and ψ within a $\pm 45^{\circ}$ window of those found in the X-ray structure with an accuracy of 94% and 91% and within a $\pm 30^{\circ}$ window of 88% and 81%.

Two methods for the assignment of motif from C^{α} coordinates are reported. For the first method motif is assigned from the dihedral angles predicted using the regression equations. If the predicted dihedral angles of a residue fall in the range of $-15^{\circ} > \phi > -90^{\circ}$ and $-10^{\circ} > \psi > -70^{\circ}$, the residue is assigned as in an α-helix; and in the range of $-90^{\circ} > \phi > -150^{\circ}$ and $95^{\circ} < \psi < 170^{\circ}$ as in a β-sheet. By the second method motif of the *i*th residue is assigned from the distance C_{i-1}^{α} to C_{i+2}^{α} (v_6) and torsional angle C_{i-1}^{α}, C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α} (v_{13}). If these values for a residue fall in the range $v_6 < 6.0$ Å and $100^{\circ} > v_{13} > 0^{\circ}$ the residue is assigned as in an α-helix. If the values are in a range $v_6 > 8.7$ Å and $|v_{13}| > 100^{\circ}$ the residue is assigned as in a β-sheet. For the twenty four proteins 23.7% of the residues by the former method and 19.6% by the latter method are assigned differently than in the PDB.

2.1. Introduction

The prediction of the backbone structure and motif of proteins from known C^{α} coordinates has attracted wide attention because these coordinates are more readily available from X-ray experiments than the coordinates of all atoms[1,2] and are often the first to be defined during crystallographic analysis. We are interested in being able to predict all the atom coordinates of the intermediate filament of the rod domain of wool protein. For this reason our attention has been directed to understanding further the relationship between the C^{α} coordinates, motif and dihedral angles of the protein backbones.

The assignment of motif is generally based on a consideration of the distances between specific C^{α} coordinates along with hydrogen bond distances or on the dihedral angles of the backbone. For example, Levitt and Greer[3] have developed three methods to assign α-helix and β-sheet. The first concerned the torsional angle (α-angle) of the four consecutive C^{α} atoms (C_{i-2}^{α} to C_{i+1}^{α}) such that if the angle is between 10° and 120° the motif of the residues i to i+3 are assigned as a right handed α-helix; if the values are between 120° and 270° the residues are assigned as a β-sheet; and for values between -90° and 0° the residues are assigned as a left handed α-helix. The second method uses hydrogen bonds to identify motif (H-bond method). The peptide nitrogen atom is assumed to be midway between C_{i-1}^{α} and C_i^{α}. The peptide oxygen atom was positioned 1 Å from the nitrogen perpendicular to the plane through the C^{α} atoms of residues i-1, i and i+1. A hydrogen bond is defined between the N_i and O_j, when $j-i > 2$, the distance between N_i and N_j is less than 6 Å and the angle between the O_i-N_i and O_j-N_j vectors is within 60°. When this is the case the residues i-1 to i+1 are considered to be in an α-helix. An algorithm was similarly developed for the assignment of β-sheet. The third method (inter C^{α}-C^{α} method) is such that if the distance between $C_i^{\alpha}$and C_{i+3}^{α} is less than 6 Å and the distance between C_i^{α} and C_{i+4}^{α} is less than 6.5 Å, the five residues i to i+4 are considered to be helical. An algorithm was similarly developed for the assignment of residues in a β-sheet.

Application of the three methods to six test proteins resulted in 20.3%, 16.8% and 24.5% of the residues being incorrectly assigned as α-helix and 63.9% 58.9% and 30.4% of the residues being incorrectly assigned as β-sheet respectively. A combination of these three methods was used to assign secondary structure of 45 globular proteins resulting in an incorrect assignment of 20% of α-helix and 52.8% of β-sheet residues. The overall error was 20.3%.

Recently, Oldfield and Hubbard,[4] reported an analysis of the relationship between C^{α} geometry and secondary structure, selecting three geometric parameters, namely θ_1 (the angle C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α}), θ_2 (the angle C_{i+1}^{α}, C_{i+2}^{α}, C_{i+3}^{α}) and τ (the torsional angle C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α}, C_{i+3}^{α}). A strong correlation between C^{α} geometry and the protein fold was found but the authors did not use the analysis to assign secondary structure.[5]

We now report a multiple regression analysis between the backbone dihedral angles and the C^{α} coordinates obtained from the X-ray crystal structures of fourteen proteins using the Statistical Analysis System (SAS) package. The non-linear regression equations have been applied to predict specific dihedral angles for each residue in the backbone of twenty-four proteins. Two methods for the assignment of secondary structural motif from C^{α} coordinates have been developed. For the first method motif is assigned by comparison of the dihedral angles predicted using the regression equations with ranges of dihedral angles previously correlated with secondary structure.[3] By the second method secondary structure is assigned by comparison of the C_{i-1}^{α} to C_{i+2}^{α} distance (v_6) and C_{i-1}^{α}, C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α} torsional angle (v_{13}) with ranges of these values previously correlated with secondary structure. The joint use of two variables (v_6 and v_{13}) makes the method different from those previously reported.

2.2. Computational methods

Computation was performed on an IBM 320 or 320H Risc 6000 computer. The program Macromodel[6] was used for protein visualisation. The PDB file of X-ray coordinates for a protein were translated to Macromodel format using MMPDB within

Macromodel. Regression analysis to relate the independent variables of four adjacent C^{α} coordinates (v_1-v_{15}) and the dependant variables (ϕ and ψ) of the *i*th residue, was carried out using the Statistical Analysis System (SAS). The following programs, written in FORTRAN, were used in the studies reported in this chapter.[7]

The program *capara* reads the atomic coordinates of the X-ray structures in Macromodel file format along with the secondary structure for the residue when there is a motif defined in the PDB data base. The program extracts the C^{α} coordinates into a separate file and calculates the dihedral angles for all the amino acids from the X-ray coordinates. There are three routines in the program; Cap1 calculates the parameters v_1-v_{15} associated with the C^{α} coordinates for all the amino acids of the protein; Cap2 selects these parameters for each of twenty amino acids; Cap3 selects the values for each amino acid in each motif region as defined in the PDB file. The output files of the program *capara* are used as input files for the statistical and regression analysis using SAS.

The program *signtest* was written to determine optimum ranges of v_6 and v_{13} for ascribing the sign of a dihedral angle. The boundaries of v_6 and v_{13} were randomly varied to find the ranges which result in the optimum assignments of sign of the dihedral angles. The program also calculates the dihedral angles defined by the X-ray structure. The program *regression* was used to calculate the values of ϕ and ψ from the regression equations and compare the values with the X-ray in different windows. The program *motif* was developed to assign the motifs of secondary structure by either of two methods. This program, like *capara,* reads the atomic coordinates of the X-ray structures in Macromodel format along with the secondary structure as defined in the PDB and extracts the C^{α} coordinates and calculates values of v_6 and v_{13}. A routine in *motif* assigns secondary structure by comparison of both v_6 and v_{13} with predetermined ranges established using the program *select.* The second routine in *motif* assigns secondary structure from the values of the dihedral angles predicted by the regression equations. The program first calculates absolute values of the dihedral angles from the regression equations and the sign for each dihedral angle is based on specific criteria for the values of v_6 and v_{13}. The secondary structure is assigned from comparison of the dihedral

angles with predetermined ranges of these values established by use of the program *signtest.* The program *motif* compares motif of a residue with that assigned in the PDB.

2.3. Results and Discussion

Twenty four proteins were used in this study and chosen because their X-ray structures have been reported with a resolution of better than 2.5 Å[8] and they represent a diverse structural variety (Table 2.1). The proteins contain a total of 3562 amino acid residues.

Table 2.1. Protein data base used in this work

Protein Name	Code	N[a]	Resol[b]	α-Helix	%[c]	β-Sheet	%[d]
Peptide Antibiotic[9]	1AMT	60	1.5	51	84	0	0
Crambin†[10]	1CRN	46	1.5	21	46	7	15
Serine Proteinase-inhibitor[11]	1CSE	337	1.2	110	33	66	20
Ribosomal Protein†[12]	1CTF	68	1.7	39	57	17	25
Electron Transport[13]	1PCY	99	1.6	5	5	65	66
Pancreatic Hormone†[14]	1PPT	36	1.37	19	53	8	19
Hydrolase(acid proteinase Zymogen)[15]	1PSG	365	1.65	10	28	147	40
Iron-Sulfur Protein†[16]	1RDG	52	1.40	0	0	0	0
COL*E Rop †[17]	1ROP	56	1.7	53	95	0	0
Isomerase[18]	1TIM	494	2.5	278	56	100	20
Steroid Binding†[19]	1UTG	70	1.34	53	76	0	0
Cytochrome C†[20]	2CCY	254	1.67	188	74	0	0
Hydrolase (O-Glycosy)[21]	2LYM	129	2.00	59	46	19	15
Hydrolase (O-Glycosy)[22]	2LYZ	129	2.00	35	27	19	15
Oxygen Binding[23]	2MHR	117	1.70	76	65	0	0
Proteinase Inhibitor (Kazal)†[24]	2OVO	56	1.50	12	21	14	25

Leucine Zipper†[25]	2ZTA	62	1.8	62	100	0	0
Flavodoxin†[26]	3FXN	138	1.8	48	35	35	25
Transferase†[27]	3CLA	125	1.75	51	41	34	27
Rat mast cell protease II[28]	3RP2	448	1.9	50	11	162	36
DNA Binding Regulatory Protein†[29]	3WRP	101	1.80	78	77	0	0
Insulin[30]	4INS	102	1.5	54	53	6	6
Trypsin Inhibitor†[31]	4PTI	58	1.5	10	17	19	33
Troponin C†[32]	4TNC	160	2.0	97	61	0	0
Total number		3562		1598		718	

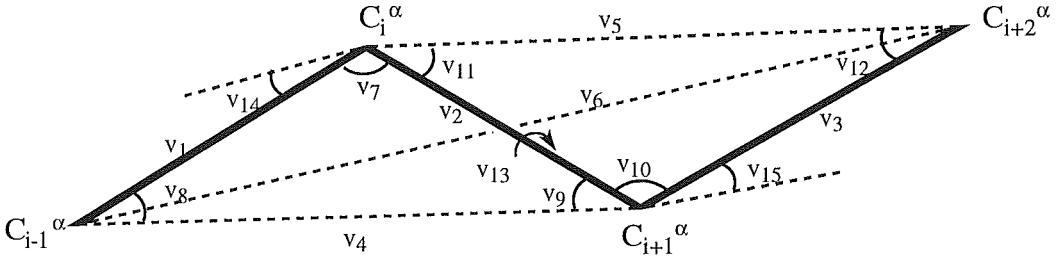
a Number of residues in the protein, b Resolution in Å, c % of the residues in α-helical motif, d % of the residues in β-sheet motif.

The atomic coordinates[33] of the twenty-four proteins were extracted from the corresponding PDB file in Macromodel format. The bond lengths and bond angles of the backbone atoms of these proteins compare favourably with previous reported values.[34]

2.3.1. Definition of C^{α} variables and dihedral angle ϕ and ψ

Since the peptide bond is almost planar and the bond lengths and angles exhibit only small deviations from mean values, atoms of a protein backbone in the *i*th peptide unit can be very closely defined[35] by the dihedral angles ϕ_i and ψ_i given the coordinates of $C_{i-1}{}^{\alpha}$,$C_i{}^{\alpha}$, $C_{i+1}{}^{\alpha}$ and $C_{i+2}{}^{\alpha}$(Figure 2.1).[36]

(a)



(b)

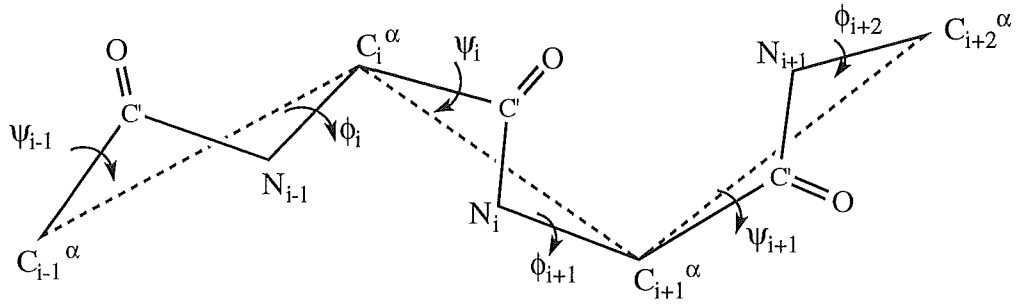


Figure 2.1 (a) Definition of variables for four adjacent C^{α} atoms, (b) definition of ϕ_i and ψ_i

2.3.2. Statistical analysis of C^{α} variables and dihedral angles

2.3.2.1. Statistical distribution

A statistical analyses of the variables (v_{1-15}) which relate four adjacent C^{α}'s (C_{i-1}^{α}, C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α}) and the dihedral angles with motif of the *i*th residue as defined in the PDB file was first carried out for fourteen proteins marked † in Table 2.1 to determine if there were any obvious correlation (see table 2.2). The values of ϕ and ψ of the amino acids in α-helical regions show normal single peak distributions and exhibit small standard deviations. The mean values of ϕ and ψ are -65° and -38° with standard deviations of 11° and 12° respectively. These values are similar to values previously reported[37] for the helical region of various selections of proteins.

Table 2.2. The mean and standard deviation of the dihedral angles and the variables v_{1-15} of the amino acid residues in the four motif regions.

	Motifs*							
	α-Helix		β-Sheet		β-turn		Random coil	
	μ	σ	μ	σ	μ	σ	μ	σ
φ	-65.09	11.12	-103.03	45.83	-61.75	60.34	-77.19	63.77

ψ	-38.82	12.35	116.46	64.77	39.36	81.67	67.16	92.74
v_1	3.82	0.042	3.79	0.10	3.79	0.09	3.79	0.07
v_2	3.82	0.042	3.79	0.08	3.79	0.09	3.79	0.07
v_3	3.82	0.04	3.79	0.04	3.79	0.09	3.79	0.09
v_4	5.48	0.17	6.55	0.47	5.91	0.58	6.25	0.63
v_5	5.51	0.24	6.52	0.50	5.90	0.57	6.23	0.64
v_6	5.30	0.58	9.62	0.92	7.44	1.65	8.37	1.57
v_7	91.94	3.87	121.95	12.57	103.6	15.35	112.75	17.76
v_8	44.03	1.95	29.02	6.311	38.25	7.82	33.65	8.95
v_9	44.04	2.00	29.03	6.321	38.18	7.61	33.59	8.85
v_{10}	92.70	5.83	120.79	14.32	103.40	15.01	112.25	17.81
v_{11}	43.64	2.93	29.62	7.139	38.33	7.65	33.89	8.99
v_{12}	43.65	2.96	29.59	7.206	38.26	7.48	33.82	8.86
v_{13}	50.38	18.05	-75.32	124.21	2.53	104.42	-27.37	112.58
v_{14}	50.10	8.41	25.95	17.70	45.63	19.15	39.27	20.32
v_{15}	50.16	8.41	25.11	18.19	45.99	20.92	39.65	22.15

*Amino acids not assigned as α-helix, β-sheet or β-turn in the PDB data base are classified for the purpose of this study as having a random coil motif. Amino acids assigned two motifs in the PDB are computed in both motifs.

In the β-sheet region the mean values of ϕ and ψ are -103° and 116° respectively, but the standard deviations of 45° and 64° are larger. In β-turn and random coil regions ϕ and ψ vary over a wide range of values. The distances between sequential C^{α} carbons, v_1, v_2 and v_3 are in the range 3.7 to 3.9 Å and have little variation with motif[38] while the distances between alternate C^{α} carbons, v_4 and v_5, are sensitive to motif.[39] The angles v_7-v_{12}, v_{14} and v_{15} are sensitive to motif and within each motif occur in a narrow range.[42] The distance from $C_{i-1}{}^{\alpha}$ to $C_{i+2}{}^{\alpha}$, v_6, is sensitive to motif with a mean value

5.31 Å (deviation 0.58 Å) in α helical regions, 9.62 Å (0.92 Å) in β-sheet, 7.44 Å (1.66 Å) in β-turn and 8.37 Å (1.58 Å) in random coil regions. The torsion angle of the four consecutive C^{α} atoms, v_{13}, has a mean value of 50° (standard deviation 18°) in α-helical regions, -75° (standard deviation 124°) in β-sheet regions, 2.5° (standard deviation 104°) in β-turn regions, and -27° (standard deviation 112°) in random coil regions.[40] The sign of the variable v_{13} reflects the folding of the C^{α} chain[41] such that if 90° > v13 > 0° the α-helix or β-turn is right handed and if -90° < v13 < 0° the α-helix or β turn is left handed.

Table 2.3. Correlation matrix of ϕ and ψ and v_1 to v_{15}

corr																
v_1	1.00															
v_2	0.07	1.00														
v_3	0.01	0.07	1.00													
v_4	-0.05	-0.14	-0.06	1.00												
v_5	-0.12	-0.04	-0.14	0.53	1.00											
v_6	-0.13	-0.13	-0.12	0.81	0.75	1.00										
v_7	-0.11	-0.22	-0.06	0.99	0.52	0.79	1.00									
v_8	0.07	0.26	0.06	-0.99	-0.51	-0.79	-0.99	1.00								
v_9	0.16	0.19	0.05	-0.99	-0.52	-0.79	-0.99	0.99	1.00							
v_{10}	-0.12	-0.11	-0.22	0.51	0.99	0.74	0.51	-0.51	-0.51	1.00						
v_{11}	0.12	0.06	0.26	-0.51	-0.99	-0.74	-0.50	0.49	0.50	-0.99	1.00					
v_{12}	0.13	0.16	0.19	-0.52	-0.99	-0.74	-0.52	0.52	0.52	-0.99	0.99	1.00				
v_{13}	0.04	0.04	0.05	-0.64	-0.27	-0.54	-0.63	0.62	0.63	-0.25	0.25	0.25	1.00			
v_{14}	0.12	0.12	0.10	-0.48	-0.41	-0.43	-0.48	0.47	0.48	-0.42	0.41	0.42	0.28	1.00		
v_{15}	0.09	0.05	0.14	-0.26	-0.53	-0.37	-0.24	0.24	0.24	-0.54	0.55	0.54	0.12	0.89	1.00	
ϕ	0.07	0.04	-0.04	-0.35	-0.13	-0.24	-0.36	0.35	0.36	-0.12	0.12	0.13	0.33	0.24	0.13	1.00
ψ	-0.15	-0.17	-0.06	0.82	0.53	0.83	0.82	-0.81	-0.82	0.52	-0.51	-0.52	-0.61	-0.38	-0.21	1.00
	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}	v_{15}	ϕ	ψ

2.3.2.2. Correlation analysis

Correlation coefficients of C^{α} variables (v_1 to v_{15}), ϕ and ψ has been calculated by using SAS. The correlation matrix shows that there are obvious correlation between the C^{α} variables and the dihedral angles (Table 2.3). For example, v_6 has a correlation coefficient of -0.26 and 0.83 with the dihedral angles ϕ and ψ respectively. The v_{13} has a correlation coefficient 0.33 and -0.61 with ϕ and ψ respectively.

2.3.2.3. Regression analysis

A mathematical relationship between the dihedral angles, ψ and ϕ, and the fifteen variables of adjacent C^{α} coordinates (v_1-v_{15}) was established by using a stepwise multiple non-linear regression programme in the Statistical Analysis System (SAS). The multiple regression equations are derived from the fourteen proteins marked † in Table 2.1. The distances are in Å and angles in radians. The most satisfactory regression equations for predicting ϕ and ψ, using adjusted R^2 as the criterion, involve the function transformations of sine and cosine and are:

$$\cos\phi = -0.87 - 0.12v_6 + 2.0\sin v_7 - 0.12\sin v_{13} \quad (2.1)$$

t-value = (-7.1) (-20.2) (18.6) (-7.6)

Adjusted R^2 = 0.63, F-value = 707, d.f = 3, 1230, SE = 0.24

$$\cos\psi = -57.1 + 0.13\sin v_6 + 29.7v_7 + 30.7\cos v_7 + 11.1\sin v_7 + 0.25\sin v_{13} \quad (2.2)$$

t-value = (-13.3) (10.4) (13.8) (15.0) (12.1) (13.6)

Adjusted R^2 = 0.87, F-value = 1700, d.f = 5, 1228, SE = 0.24

Where R is the correlation coefficient of the regression equation, F-value is the value of F-test of the regression equation, t-value is the value of t-test of each term in regression equation, d.f is the degree of freedom, SE is the sum of squared residues.

These equations are surprisingly reliable for predicting the absolute value of the dihedral angles of the protein backbone from the C^{α} coordinates, affording R^2 values of 0.63 for cosϕ and 0.87 for cosψ respectively. When the equations are determined for the twenty-four proteins the values of t, R^2, F, d.f and the regression coefficients were similar to those reported above.[42] Only the absolute values of the dihedral angles ϕ, ψ can be obtained using the inverse function of cosine from the regression equations (2.1) and (2.2) and the sign must be established by alternative means. The variables v_6 and v_{13} have been used for this purpose. A positive sign is assigned for ϕ when 6.0 Å $\geq v_6 >$ 4.0 Å and -3.4° $\geq v_{13} >$ -83.7° and ψ when 13.0 Å $\geq v_6 >$ 7.5 Å and $v_{13} \geq$ 63.5° or < -97.2°; 5.7 Å $\geq v_6 >$ 4.0 Å and-17.2° $\geq v_{13} >$ -97.2°. A negative sign was assigned when v_6 and/or v_{13} lie outside these ranges. The ranges were established using the program *signtest* by random variation of the boundaries of v_6 and v_{13} to allow optimum assignment to the sign of the dihedral angles for the fourteen protein sample. Using these criteria for the fourteen proteins (1234 values of ϕ and of ψ), 95% of the values of ϕ and 94% of ψ are predicted with the sign of the dihedral angle as defined in the PDB. The method is called the Non-linear Regression Distance Torsion (NLRDT) method.

2.3.3. Prediction of dihedral angles from C^{α} coordinates

The success of this method in predicting dihedral angles is measured by the percentage of calculated dihedral angles which fall within a window of ± 30° or ± 45° of the value determined by the X-ray analysis.[43] For the twenty-four proteins 94% of the values of ϕ and 91% of the values of ψ are predicted within a ± 45° window and 88% of the values of ϕ and 81% of the values of ψ are predicted within a ± 30° window (see Table 2.4).

Table 2.4. Percentage of predicted values of dihedral angles within windows of ± 30° and ± 45° of the values in the X-ray structures

		ϕ				ψ			
Protein	C^{α}	$\pm 30^{\circ}$		$\pm 45^{\circ}$		$\pm 30^{\circ}$		$\pm 45^{\circ}$	
	Number	Out	A%	Out	A%	Out	A%	Out	A%
1AMT	51	0	100	0	100	1	98	0	100
1CRN	43	6	86	0	100	8	81	4	91
1CSE	331	59	82	28	92	83	75	46	86
1CTF	65	8	88	3	95	16	75	8	88
1PCY	96	16	83	8	92	35	64	15	84
1PPT	33	3	91	1	97	5	85	2	94
1PSG	359	60	83	33	91	86	76	45	88
1RDG	49	10	80	5	90	18	63	8	84
1ROP	53	1	98	1	98	3	94	1	98
1TIM	488	109	78	63	87	114	77	61	88
1UTG	67	4	94	1	98	8	88	4	94
2CCY	248	21	92	10	96	31	88	11	96
2LYM	126	23	82	16	87	38	70	20	84
2LYZ	126	30	76	17	87	40	68	15	88
2MHR	114	13	89	6	95	14	88	6	95
2OVO	53	8	85	3	94	9	79	7	87
2ZTA	56	0	100	0	100	0	100	0	100
3FXN	135	21	84	11	92	30	78	14	90
3CLA	122	13	89	5	96	13	89	3	98
3RP2	442	111	75	58	87	125	72	62	86
3WRP	98	8	92	5	95	9	91	5	95
4INS	90	5	94	3	97	13	86	6	93
4PTI	55	8	86	4	93	15	73	10	82
4TNC	157	10	94	8	95	23	85	12	92
Overall			88		94		81		91

A%-The percentage of predicted values within the window. "Out" The number of predicted values outside the window.

Table 2.5. Percentage of predicted values of dihedral angles for the different amino acids in twenty-four proteins within windows of ± 30° and ± 45° of the values in the X-ray structures

			ϕ				ψ		
Amino	Number	± 30°		± 45°		± 30°		± 45°	
acid		Out	A%	Out	A%	Out	A%	Out	A%
Ala	322	48	85	19	94	55	82	28	91
Arg	121	17	86	8	93	25	79	11	90
Asn	144	36	75	21	85	37	74	21	85
Asp	199	25	87	14	92	49	75	27	86
Cys	89	14	84	4	95	23	74	16	82
Gln	104	7	93	4	96	13	87	9	91
Glu	237	20	91	10	95	30	87	11	95
Gly	291	118	59	105	64	114	60	56	80
His	72	17	76	9	87	13	82	5	93
Ile	186	18	90	2	99	26	86	12	93
Leu	274	21	92	9	96	45	83	24	91
Lys	220	29	86	18	91	42	80	19	91
Met	69	6	91	1	98	8	88	6	91
Phe	125	15	88	9	92	25	80	13	89
Pro	151	27	82	7	95	47	68	29	80
Ser	228	47	79	21	90	58	74	27	88
Thr	180	28	84	8	95	43	76	21	88
Trp	50	4	92	0	100	15	70	7	86

Tyr	114	21	81	7	94	34	70	13	88
Val	260	29	89	13	95	37	85	12	95
Overall	3436	522	85	93	97	612	82	244	93

Twenty-one amino acids are not defined in the PDB file. A%-The percentage of amino acids predicted correctly. "Out"-the number of amino acids predicted outside the window.

The accuracy for prediction of ϕ and ψ in the proteins with high α-helical motif content, such as 2ZTA, 1AMT, 1ROP, 2CCY, 4TNC, is excellent. For example for the protein 2ZTA which has an α-helical coiled-coil two chain structure the accuracy of prediction for both ϕ and ψ is 100% in both window frames. Not surprisingly, in proteins with a more irregular structure which contain regions of β-turn and random coil motif and for the larger proteins, 1CSE, 1PSG, 1TIM and 3RP2 with complicated secondary structures the accuracy of prediction is lower (Table 2.4).

The regression equations have been applied[44] to the four secondary structure motif regions of the twenty-four proteins and to individual amino acids. In the regular α-helical regions, the percentage of correctly predicted dihedral angles is high; 95% and 91% of the predicted values of ϕ and ψ fall within a $\pm 30^{\circ}$ window respectively and 98% and 96% of the values of ϕ and ψ fall within a $\pm 45^{\circ}$ window respectively. In the β-sheet region the percentage of dihedral angles ϕ and ψ predicted within the window is 88% and 80% within a $\pm 30^{\circ}$ window and 96% and 91% within a $\pm 45^{\circ}$ window respectively. In the β-turn motif the success in predicting values of ϕ and ψ within a $\pm 30^{\circ}$ window is 76% and 63% and 86% and 84% within a $\pm 45^{\circ}$ window respectively. In the random coil region, the success in predicting values of ϕ and ψ within a $\pm 30^{\circ}$ window is 72% and 67% and 85% and 83% within a $\pm 45^{\circ}$ window. For individual amino acids the percentage of correctly predicted dihedral angles is shown in Table 2.5.

Overall, the average percentage of correctly predicted dihedral angle for twenty amino acids within a $\pm 30^{\circ}$ window is 85% and 82% and within a $\pm 45^{\circ}$ window 97%

and 93% respectively. The accuracy for predicting dihedral angles of the non-polar amino acid residues, alanine, isoleucine, leucine, methionine, phenylalanine and valine are above the average and the polar amino acids, asparagine, asparagic acid, glutamine, glutamic acid, cystine, serine and threonine, are below the average. Predictions for proline are modest and for glycine poorest because the absence of a side-chain allows the amino acid to adopt a wide range of conformations.

These values of the dihedral angles for each amino acid of the protein obtained in this way cannot however in themselves be used to define coordinates of the backbone atoms since they result in unacceptable bond lengths and angles of the backbone atoms in relation to the defined C^{α} coordinates. The study shows, that the dihedral angles of the protein have a high probability of occurring within a window of $\pm$ 30° of the value computed from the NLRDT method. This allows for angles for each residue of the protein to be chosen by Monte Carlo methods from within a $\pm$ 30° window of the value determined from the regression equations and which meet the criteria of positioning the backbone atoms with acceptable bond lengths and angles (see chapter 3).

2.3.4. Assignment of secondary structure from C^{α} coordinates

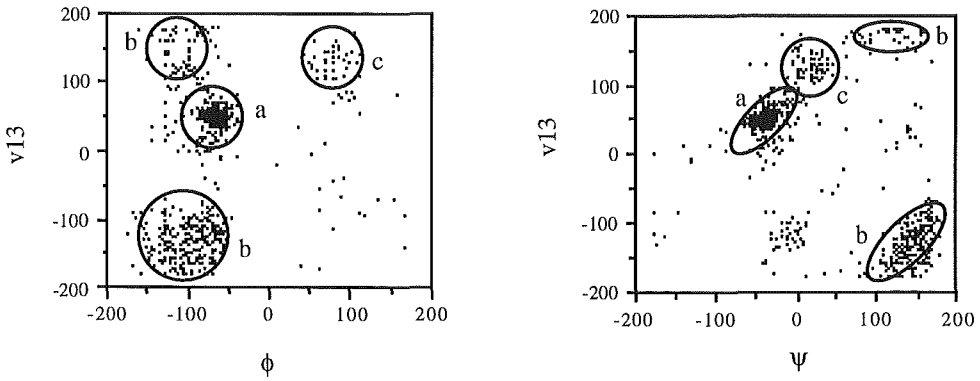
2.3.4.1. Regression dihedral angle method (method 1)

We report two methods for the assignment of motif from C^{α} coordinates. The first is from a comparison of the dihedral angles of an *i*th residue, predicted by the non-linear regression equations relating coordinates of C_{i-1}^{α} to C_{i+2}^{α}, and predetermined ranges of these values. The distribution of dihedral angles ϕ and ψ is normally in a range of -60° $\pm$ 30° and -40° $\pm$ 30° for an α-helix and the range -90° $\pm$ 30° and 120° $\pm$ 30° for a β-sheet.[3] In order to define specific ranges for ϕ and ψ that can best be used as a criterion for assigning secondary structure, boundaries for ϕ and ψ were randomly extended at both ends by up to 30° for ϕ and also for ψ by using the program *select.* The ranges chosen were those which result in the optimum number of correct assignments of motif for the sample proteins. From the analysis with the fourteen protein sample the most satisfactory

ranges for assigning α-helical secondary structure are: -15° to -90° of ϕ and -10° to -70° of ψ; for β-sheet -80° to -150° of ϕ and 90° to 170° of ψ. A residue with values of ϕ and ψ not in these ranges is assigned as a random coil motif for convenience. The method is referred to as the *RDA* method (Regression Dihedral Angle method). Left-handed and right handed α-helix and β-turns are not specifically designated but can be distinguished from values of v_{13}.

2.3.4.2. Distance and torsion of C^{α} method (method 2)

The second method of assigning secondary structure to the *i*th amino acid is by comparison of the distance C_{i-1}^{α} to C_{i+2}^{α} (v_6) and torsional angle C_{i-1}^{α}, C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α} (v_{13}) with predetermined ranges in these values. The method is referred to as the *DTC* method (Distance and Torsion of C^{α}). The variables v_6 and the v_{13}, are the most sensitive to motif and range for values between 4.0 Å and 14 Å and between -180° and 180° respectively. The plots of v_6 and v_{13} versus ϕ or ψ (Figure 2.2) confirm a relationship between the dihedral angle and these two variables. This relationship could have the potential to predict motif. The motif regions of these plots are identified from the mean value and distribution of the variables shown in Table 2.3.



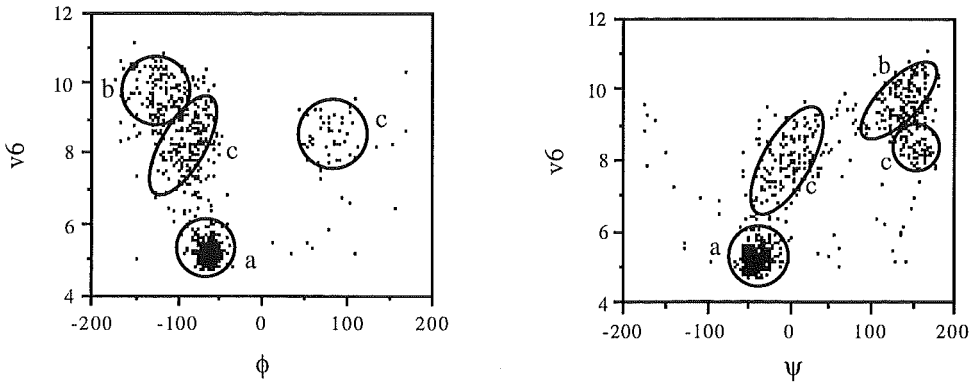


Figure 2.2. Plot of ϕ, ψ vs v_{13} and v_6 (a, α-helix region; b, β-sheet region; c, β-turn or random coil region)

A statistical analysis for fourteen proteins of the parameters which describe four adjacent C^{α}'s and motif has shown that v_6 and v_{13} correlate best with motif. For example residues in an α-helix have a range of 4.5 Å < v_6 < 6 Å and 25° < v_{13} < 75° and in a β-sheet a range 8 Å < v_6 < 14 Å and either -180° < v_{13} < -100° or 150° < v_{13} < 180°. When values of v_6 and v_{13} occur in these ranges respectively the α-helix and β-sheet motifs can be assigned. The relationship of the values of v_6 and v_{13} and motif have been extensively examined empirically using the program *select* for fourteen proteins in order to define specific ranges for these variables that can best be used as a criterion for assigning secondary structure.

The boundaries of v_6 and v_{13} are randomly varied within a range of 1 Å of v_6 and 40° of v_{13}. The most satisfactory ranges to assigning secondary structure are: α-helix; 4.0 Å < v_6 < 6.0 Å and 0° < v_{13} < 100°; β-sheet; v_6 > 8.7 Å and | v_{13} | > 100°; β-turn, 6.0 Å < v_6 < 8.7 and-100° < v_{13} < 100° .[45] If the values of v_6 and v_{13} of a residue fall in the range of v_6 < 6 Å and 0° < v_{13} < 100°, the residue is assigned as an α-helix. If the values of v_6 and v_{13} of a residue fall to the range of v_6 > 8.7 Å and v_{13} < -100° or v_{13} > 100°, the residue is assigned as a β-sheet. A residue not assigned α-helix and β-sheet motif is assigned as a random coil.

Table 2.6. Comparison of the methods for assigning secondary structure

		α-helix			β-sheet	
Protein	X-ray†	*DTC*	*RDA*	X-ray	*DTC*	*RDA*
1CRN	7-19	7-18	6-19	2-4	2-3	2-5
	23-30	23-29	23-29	32-35	32-34	32-37
1CTF	13-25	13-23	12-23	2-8	2-4	2-5
	27-34	28-36	28-36	41-46	-	39-41
	47-58	48-60	48-61	64-66	63-66	64-66
1PPT	14-32	14-31	14-31	1-8	3-7	-
2LYM	4-16	5-14	5-14	1-3	1-3	1-3
	24-36	25-34	25-34	43-46	42-45	43-45
	79-81	80-83	80-83	50-54	50-52	50-55
	88-101	89-99	89-99			
	108-115	109-113	108-115			
	119-124	120-125	120-123			
2OVO	33-42	34-43	32-43	-	4-6, 15-16	4-6, 13-16
				22-25	22-24	22-23
				29-32	-	-
				50-54	-	48-49
						53-54
3FXN	11-25	11-25	11-26	1-5	1-6	1-8
	64-73	66-72	40-44	29-34	28-33	28-33
	93-104	94-105	67-74	49-53	48-55	51-56
	125-135	125-136	94-106	81-88	81-88	80-88
			125-136	109-119	115-117	115-117

3CLA	14-23	13-20	13-20	3-5	-	2-3
	37-44	37-44	37-44	26-35	25-33	27-33
	50-68	50-63	50-63	70-73	71-72	70-71
	108-121	66-68	63-68	76-80	77-78	77-80
		108-123	108-123	85-90	82-90	82-90
				97-101	96-101	96-104
4PTI	-	3-6	3-6	16-25	18-21	20-23
	47-56	48-54	47-55	28-36	29-34	29-33
						44-46
Total	259			153		
‡Difference		47	53		75	92
(%)		18.1	20.4		49.0	61.7

† References are given in Table 2.1. The column lists the residue number.‡ Difference means an incorrect assignment compared with the X-ray structure.

2.3.5. Comparison of assignments with the X-ray structure

The *RDA* and *DTC* methods have been separately applied to assign secondary structure to the twenty four proteins listed in table 2.1. We selected eight small and simple proteins which have residues assigned as α-helix and β-sheet in the PDB as an example to display the detail of the assignments (Table 2.6). The proteins are reported in the PDB to have a total of 23 α-helical segments and 25 β-sheet segments. The number of residues in the eight proteins is 656 of which 259 are in α-helix and 153 residues are in β-sheet. The secondary structure assigned using these two methods is compared with that of the PDB in Table 2.6.[46]

The *RDA* method finds all the reported α-helical segments and an additional helical segment in each of 3FXN and 3CLA (Table 2.6). Of the 259 residues defined as α-helix

in the PDB, 53 residues (20.4%) were not correctly assigned by this method. The method missed a β-sheet segment in each of 1PPT and 2OVO, and the four additional β-sheet segments are found in 2OVO and 4PTI. This method incorrectly assigns 92 of the 153 β-sheet residues (60.1%). More additional α-helical and β-sheets segments were determined by this method than by the *DTC* method.

The *DTC* method found all 23 reported α-helical segments and an additional helical segment in 3CLA. The method assigns 47 of the 259 residues in a α-helix differently from the PDB (18.1%). The method found 23 segments of β-sheet motif, of which 21 segments are reported in the PDB. Four β-sheet segments are missed in 1CTF, 2OVO and 3CLA and two extra β-sheet segments are found in 2OVO. The method defines 75 of the 153 residues as β-sheet (49.0%). The residues at the ends of α-helical and β-sheet segments are occasionally missed by the method. The assignment of α-helix and β-sheet by the *DTC* method is closer to the PDB assignment than by the *RDA* method. Assignment of β-sheet is more difficult than for α-helix and the assignments agree much more closely with the PDB for α-helices than β-sheets by both methods.

Table 2.7. Comparison of predicted secondary structure with the PDB

	DTC method					*RDA* method				
Protein	α-Helix		β-Sheet			α-Helix		β-Sheet		
	E[a]	M[b]	E	M	d%	E	M	E	M	d%
1AMT	0	0	0	0	0	0	0	0	0	0
1CRN	0	2	0	2	8	1	1	3	0	10
1CSE	17	16	25	31	26	14	11	47	29	29
1CTF	4	4	1	10	27	6	4	2	9	30
1PCY	0	2	1	22	25	3	2	1	23	29
1PPT	1	1	1	5	22	1	1	4	0	13
1PSG	18	20	35	43	31	30	18	47	50	39
1RDG	0	0	3	0	5	3	0	6	0	17
1ROP	0	4	0	0	7	0	4	0	0	7

1TIM	0	65	33	22	28	5	49	43	30	25
1UTG	1	2	0	0	4	1	2	1	0	5
2CCY	14	6	1	0	8	20	5	2	0	10
2LYM	3	14	1	6	18	2	15	3	4	18
2LYZ	14	2	0	7	17	16	2	4	6	21
2MHR	6	0	0	0	5	6	0	6	0	10
2OVO	1	1	4	4	16	2	0	9	1	17
2ZTA	0	0	0	0	0	0	0	0	0	0
3FXN	2	4	2	10	15	10	4	8	11	23
3CLA	3	5	4	10	17	6	0	7	10	18
3RP2	4	10	42	42	21	24	4	56	48	29
3WRP	1	3	0	0	3	3	2	1	0	5
4INS	5	8	20	0	32	2	12	22	1	36
4PTI	0	3	0	9	20	0	1	5	11	29
4TNC	14	10	8	0	20	22	6	6	0	21
Total	107	182	180	225		176	143	279	240	

a extra number of residues, b missed number of residues, d % difference of assignment compared with the X-ray structure.

The 'extra residues' and the 'missed residues' in the assignment of secondary structure in twenty four proteins compared with the assignment reported in the X-ray are given in the table 2.7. In the *RDA* method, there are 176 'extra residues' and 143 'missed residues' of α-helix motif, 279 'extra residues' and 248 'missed residues' of β-sheet motif compared with the X-ray structure. Overall 846 of the 3563 residues (23.7%) are 'missed' or 'extra residues'. 319 of the 1598 α-helical residues (19.9%) and 527 of the 718 β-sheet residues (73.4%). For the *DTC* method, there are 107 'extra residues' and 182 'missed residues' of α-helix motif and 180 'extra residues' and 225 'missed

residues' of β-sheet motif. Overall there are 287 'extra residues' and 407 'missed residues', a total 'difference' of 700 of the 3563 residues (19.6%). When these differences are calculated separately for α-helices and β-sheets, 289 residues of the 1598 α-helical residues (18.1%) and 407 of the 718 β-sheet residues (56.7%) are 'extra' or 'missed'. The number of 'missed residues' is greater than the 'extra residues' for α-helical regions by both methods. For β-sheet regions the number of 'missed residues' is greater than the 'extra residues' by the *DTC* method and is less than the 'extra residues' by the *RDA* method for the proteins examined.

2.3.6. Comparison of assignments with previous methods

Both the methods are remarkably reliable in the assignment of secondary structure rich of α-helix but less satisfactory for the assignment of β-sheet. The secondary structure in a large protein is always more complicated to assign and the results are less reliable. For the *RDA* method 23.7% (α-helix 19.8%) of the total number of residues are incorrectly assigned and 19.6% of residues (α-helix 18.1%) are incorrectly assigned by the *DTC* method. The Levitt/Greer method results in 20.3% (α-helix 20.3%) of residues being incorrectly assigned. For the *RDA* method, in the case of β-sheets, 73.4% of residues are incorrectly assigned and 56.7% are incorrectly assigned by the *DTC* method compared with 53.0% by the Levitt/Greer (L/G) method.

We have chosen Trypsin inhibitor (4PTI) to compare the various methods for assigning secondary structure (Table 2.8). Two segments were assigned which have not been assigned in the PDB file namely the H1 segment (residues 3 to 6) assigned by both the *DTC* and the *RDA* method to be an α-helical segment and segment E3 (residue 44 to 46) assigned by the *RDA* method as a β-sheets. These assignments agree with the Levitt/Greer (L/G) method and the KS method. The KS method[47] uses the coordinates of all the backbone atoms to assign secondary structure and not just the C^{α} carbons. Inspection of the X-ray structure of PTI in Macromodel shows residues 3-6 of H1 region are in an α-helix and residues 44-46 of E3 region are in a β-sheet.

Table 2.8. Comparison of methods for assigning secondary structure to PTI.

		Method					
Motif	Reported in X-ray	α-angle (L/G)	H-bond (L/G)	Inter-Cα-Cα (L/G)	H-bond (K/S)	This work *DTC*	This work *RDA*
H1	-b	3-6	2-7	-b	3-6	3-6	3-6
E1	16-25	16-24	14-25	16-18,23-25	18-24	18-21	20-23
E2	28-36	30-36	28-37	28-30,35-37	29-35	29-34	29-33
E3	-b	-b	43-46	-b	45-45	-b	44-46
H2	47-56	48-56	47-55	47-56	48-55	48-54	47-55

H = α-helix, E = β-sheet. b = no secondary structure assigned.

2.3.7. Discussion

The two methods developed in this chapter assign α-helix motif well and for the proteins 1AMT, 1UTG, 2CCY, 2MHR, 2ZTA and 3WRP which are rich in α-helices more than 90% of residues are correctly assigned by both methods. In proteins rich in β-sheet, the assignment of secondary structure is less accurate. One of the reasons for the difference between the assignments made by these methods from those in the PDB is that secondary structure assignments in the data base are often incomplete and in some cases incorrect. For example, in protein 1CSE, residues 63-74 are assigned as α-helix, but the dihedral angles ϕ and ψ of the alanine residues 72 and 73 are -133.6°, 16.4° and -55.8°, 138.6° respectively which is outside the range assigned as α-helical. In protein 3RP2, there are two sub-units which each contain 228 residues. The first segment of α-helix was assigned from residue 164 to 168, but the dihedral angles (ϕ, ψ) within the sequence (164 F, -152.1°, 34.8°; 165 Q, -136.9°, 146.5°; 166 V, -109.9°, 141.8°; 167 C, -90.9°, 131.2°; 168 V, -134.3°, 128.6°) are outside those associated with α-helical motif. For the

twenty four proteins 8.5% of the residues assigned in the PDB as α-helix have the dihedral angle outside the normal range (ϕ -60° ± 30° and ψ -40° ± 30°) defining the α-helix configuration and 12.0% of the residues defined in the PDB as β-sheet are outside the values defining a β-sheet (ϕ -90° ± 30° and ψ 120° ± 30°).

The *DTC* method is overall better than *RDA* method in assigning secondary structure. The former method requires only the coordinates of four adjacent C^{α} atoms and is sensitive to the handiness of the C^{α} chain. v_{13} can be used to distinguish left and right handed α-helices and β-turns. The weakness of this method is that it does not always predict the ends of an α-helix or a β-sheet precisely and does not recognise short β-sheets segments well. The *RDA* method does recognise short β-sheet segments and has the advantage of predicting dihedral angles of the backbone, but is not as accurate as the *DTC* method in assigning motif. The dihedral angles obtained from the regression equations will sometimes be inappropriate because they lead to unacceptable bond lengths and angles between N_i and C_i^{α}.[48] The *RDA* method is somewhat less accurate than the *DTC* method.

2.4. Conclusion

Non-linear regression equations have been developed to establish a relationship between the C^{α} coordinates of a protein and the backbone dihedral angles ϕ and ψ. This relationship has been used to predict absolute values of the dihedral angle ϕ and ψ of a protein backbone knowing only the C^{α} coordinates of the protein. The sign of the dihedral angles ϕ_i and ψ_i of the *i*th amino acid is assigned from a comparison of the C_{i-1}^{α} to C_{i+2}^{α} distance and the C_{i-1}^{α}, C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α} torsional angle with predetermined ranges of these values established to best correlate with sign. For α-helical regions, 98%, 96% and 95%, 91% respectively of the dihedral angles ϕ and ψ fall within a ± 45° and a ± 30° windows of the value in the X-ray structure. For β-sheet regions 96% and 91% fall within a ± 45° window and 88% and 81% within a ± 30° window. The overall accuracy for the prediction of the backbone dihedral angles for the twenty-four proteins is 94% and 91% respectively within a ± 45° window and 88% and 81% within ± 30° window. The

NLRDT method is most successful in predicting dihedral angles of proteins rich in α-helix and β-sheet. The use of this methodology to build models of coiled-coil proteins from C^{α} coordinates therefore offers potential. The secondary structure motif of proteins can be assigned by either the *RDA* or *DTC* methods. The latter method is simpler and more accurate. Both methods are somewhat better than previously reported methods and are, like all methods, most successful for the assignment of secondary structure for α-helices and β-sheet. We will report the application of the former method to coiled-coil proteins in chapter three.

References

1 Correa P. E. (1990) Proteins **7,** 366-377. Jones T. A., Thirup S. (1986) *EMBO J.* **5,** 819-822. Holm L., Sander C. (1991) *J. Mol. Biol.* **218,** 183-194.

2 Reid L. S., Thornton J. M. (1989) *Proteins* **5,** 170-182.

3 Levitt M., Greer J. (1977) *J. Mol. Biol.* **114,** 181-293.

4 Oldfield T. J. Hubbard R. E. (1994) *Protein* **18,** 324-337.

5 Wilmot C. M., Thornton J. M. (1988) *J. Mol. Biol.* **203,** 221-232. Lewis P. N., Momany F. A., Scheraga H. A. (1973) *Biochem. Biophysics. Acta* **303,** 211-229.

6 Mohamadi F., Richards N. G. J., Guida W. C., Liskamp R., Lipton M., Caufield C., Chang G., Hendrickson T., Still W. C. (1990) *J. Computational Chemistry* **11,** 440-467.

7 Available from Kaiwan Gan, Chemistry Department, University of Canterbury.

8 Lesk A. M. Protein Architecture (1991) Oxford University Press, Oxford New York Tokyo p15.

9 Fox Jr. R. O., Richards F. M. (1982) *Nature* **300,** 325-330.

10 Teeter M. M., Hendrickson W. A. (1979) *J. Mol. Biol.* **127,** 219-223. Teeter M. M., Hendrickson W. A. (1981) *Biochemistry* **20,** 5437-5443.

11 Bode W., Papamokos E., Musil D. (1987) *Eur. J. Biochem.* **166,** 673-692.

12 Leijonmarck M., Liljas A. (1987) *J. Mol. Biol.* **195,** 555-580.

13 Guss J. M., Freeman H. C. (1983) *J. Mol. Biol.* **169,** 521-563.

14 Blundell T. L., Pitts J. E., Tickle I. J., Wood S. P., Wu C. W. (1981) *Proc. Nat. ACAD. SCI. U.S.A.* **78,** 4175-4179.

15 Rao S. N., Koszelak S. N., Hartsuck J. A. (1977) *J. Biol. Chem.* **252,** 8728-8730.

16 Frey M., Sieker L., Payan F., Haser R., Bruschi M., Pepe G., Le-Gall J. (1987) *J. Mol. Biol.* **197,** 525-541.

17 Banner D W., Michael K., Tsernoglou D. (1987) *J. Mol. Biol.* **196,** 657-675.

18 Banner D. W., Bloomer A. C., Petsko G. A., Phillips D. C., Wilson I. A. (1976) *Biochem. Biophys. Res. Comm.* **72,** 146-155.

19 Morize I., Surcouf E., Vaney M. C., Epelboin Y., Buehner M., Fridlansky F., Milgrom E., Mornon J. P. (1987) *J. Mol. Biol.* **194,** 725-739.

20 Finzel B. C., Weber P. C., Harpman K. D., Salemme F. R. (1985) *J. Mol. Biol.* **186,** 627-643.

21 Kundrot C. E., Richards F. M. (1987) *J. Mol. Biol.* **193,** 157-170.

22 Diamond R. (1974) *J. Mol. Biol.* **82,** 371-391.

23 Sherife S., Hendrickson W. A., Smith J. L. (1987) *J. Mol. Biol.* **197,** 273-296.

24 Empie M. W., Laskowski M. Jr. (1982) *Biochemistry* **21**, 2274-2284.

25 O'Shea E. K., Klemm J. D., Kim P. S., Alber T. (1991) *Science* **254,** 539-544.

26 Smith W. W., Burnett R. M., Darling G. D., Ludwig M. L. (1977) *J. Mol. Biol.* **117,** 195-225.

27 Leslie A. G. W. (1990) *J. Mol. Biol.* **213,** 167-186.

28 Remington S. J., Woodbury R. G., Reynolds R. A., Matthews B. W., Neurath H. (1988) *Biochemistry* **27,** 8097-8105.

29 Zhang R. G., Joachimiak A., Lawson C. L., Schevitz R. W., Otwinowski Z., Sigler P. B. (1987) *Nature* **327,** 591-597.

30 Bordas J., G. Dodson G., Grewe H., Koch M. H. J., Krebs B., Randall J. (1983) *Proc. R. Soc. London Ser.* B **219,** 21-39.

31 Marquart M., Walter J., Deisenhofer J., Bode W., Huber R. (1983) *Acta Crystallogr.* Sect B. **39,** 480-490.

32 Sundaralingam M., Bergstrom R., Strasburg G., Rao S. t., Roychowdhury P., Greaser M., Wang B.C. (1985) *Science* **227,** 945-948.

33 Hydrogen atoms are not included.

34 Engh R. A., Huber R. (1991) *Acta Cryst.* **A47,** 392-400.

35 IUPAC-IUB Commission on Biochemical Nomenclature, Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. (1970) *Biochemistry* **9,** 3471-3479.

36 v_1 is the distance C_{i-1}^{α}, C_i^{α}. v_2 is the distance C_i^{α}, C_{i+1}^{α}. v_3 is the distance C_{i+1}^{α}, C_{i+2}^{α}. v_4 is the distance C_{i-1}^{α}, C_{i+1}^{α}. v_5 is the distance C_i^{α}, C_{i+2}^{α}. v_6 is the distance C_{i-1}^{α}, C_{i+2}^{α}. v_7 is the angle C_{i-1}^{α}, C_i^{α},C_{i+1}^{α}. v_8 is the angle C_i^{α}, C_{i-1}^{α}, C_{i+1}^{α}. v_9 is the angle C_i^{α}, C_{i+1}^{α}, C_{i-1}^{α}. v_{10} is the angle C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α}. v_{11} is the angle C_{i+1}^{α}, C_i^{α}, C_{i+2}^{α}. v_{12} is the angle C_{i+1}^{α}, C_{i+2}^{α}, C_{i-1}^{α}. v_{13} is the torsional angle C_{i-1}^{α}, C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α}. v_{14} is the angle of the line C_{i-1}^{α}, C_i^{α} to the plane C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α}. v_{15} is the angle of the line C_{i+1}^{α}, C_{i+2}^{α}, to the plane C_{i-1}^{α}, C_i^{α}, C_{i+1}^{α}.

37 Levitt M. (1983) *J. Mol. Biol.* **170,** 723-764. Presta L. G., Rose G. D. (1988) *Science* **240,** 1632-1641.

38 Roterman I. K., Lambert M. H., Gibson K. D., Scheraga H. A. (1989) *J. of Biomolecular Structure & Dynamics* **7,** 421-453.

39 In α-helical regions the mean values of v_4 and v_5 are 5.48 Å and 5.51 Å respectively and the standard deviations are 0.17 Å and 0.24 Å. In β-sheet regions the mean values are 6.54 Å and 6.52 Å with standard deviations of 0.47 Å and 0.50 Å respectively. Their standard deviations along with those of the dihedral angles ϕ, ψ increase in the order: α-helix < β-sheet < β-turn < random coil.

40 There are at least two peaks in the distribution of v_{13} in the motif of β-sheet, β-turn and random coil.

41 The 'torsional' angle of four consecutive C^{α} atoms $C_{i-1}{}^{\alpha}$ to $C_{i+2}{}^{\alpha}$ is defined as positive (0° to 180°) when there is a clockwise twist in the direction from $C_{i-1}{}^{\alpha}$ to $C_{i+2}{}^{\alpha}$ along the $C_i{}^{\alpha}$ to $C_{i+1}{}^{\alpha}$ axis and negative for an anticlockwise twist (0° to -180°).

42 As a check of the program residuals are calculated with SAS as well as *Regression* and both programs gave the sum of residues = zero for the fourteen-protein sample and the sum of squared residues is the same for each program.

43 Levitt M. (1992) *J. Mol. Biol.* **226,** 507-533.

44 The accuracy for prediction of ϕ and ψ in both windows is high and not significantly dependent of the size of the protein set.

45 The prediction starts from the second residue and continues till the last two residues of the protein are reached.

46 Not all amino acid residues are assigned motif in the PDB file and for the twenty-four proteins in this study some 50% have not been assigned motif.

47 Kabsch W., Sander C. (1983) *Biopolymers* **22,** 2577-2637.

48 Dihedral angles of the protein have a high probability of occurring within a window of ± 30° of the computed value, and angles from such ranges can be selected for each residue of the protein by Monte Carlo methods. In accepting the selected values criteria which allow for the positioning the backbone atoms with appropriate bond lengths and angles must be met.

Chapter Three

The Reconstruction of a Protein Backbone from C^{α} Coordinates

Summary

A Monte Carlo Protein Building (MCPB) method to construct the backbone and C^{β} atomic coordinates of twenty-four proteins from known C^{α} coordinates is reported. The method selects values of dihedral angles from either ± 30° windows of the dihedral angle calculated for that amino acid by the NLRDT method, or from ranges established from a statistical analysis of the relationship between dihedral angles of the backbone and C^{α} coordinates for a protein data base. The averaged coordinates from ten backbone models of a protein were used to define a mean structure that was refined by energy minimisation using the AMBER force field (GB/SA). By the latter method the average atomic deviation and r.m.s.d. of the backbone and C^{β} atoms is between 0.14 Å and 0.32 Å (average 0.22 Å) and 0.22 Å and 0.61 Å (average 0.43 Å) respectively. A comparison with other methods is made.

3.1. Introduction

In recent years the construction of a full-atom representation of a protein from C^{α} coordinates has received considerable attention. Commonly employed procedures[1] make use of "backbone dictionaries" and are referred to as homology building methods. The protein sequence and C^{α} coordinates of a target protein are compared with a database of crystal structure coordinates for segments of the same sequence and similar topology. The all atom atomic coordinates of the residues in the analogous X-ray crystal structures are transposed to the target protein. The method has limitations in regions where no similar patterns of sequence and topology exist in the data base, and in the linking regions of segments of known topology.

Several methods that avoid these problems have been developed including those of Purisima and Scheraga,[2] Reid and Thornton,[3] Correa,[4] Holm and Sander,[5] Jones and Thirup,[6] and Rey and Skolnick.[7] For example Correa constructed the protein sequentially from C^{α} coordinates with energy minimisation after positioning each backbone and β-carbon atom, but this is expensive in computer time even for medium sized proteins. More recently, Mathiowetz and Goddard[8] have reported a Dihedral Probability Grid Monte Carlo (DPG-MC) method to build protein models from C^{α} coordinates. The dihedral angles of the backbone and the torsional angles of the side-chains are randomly selected for each residue from predetermined ranges in a probability matrix. The probabilities were assigned to the protein backbone and side-chain dihedrals according to their distributions in known protein structures. Mandal and Linthicum[9] have described a modelling algorithm, PROGEN, which uses an optimal geometry parameter database established by examining the statistical correlation among twenty-three different intra- and inter- peptide geometric parameters relating C^{α} distances for each amino acid in a library of nineteen proteins from the PDB. The initial model is refined by energy minimisation and molecular dynamics techniques while keeping the C^{α} atoms fixed.

We now report a Monte Carlo Protein Building (MCPB) method to reconstruct the backbone and C^{β} atomic coordinates of twenty-four proteins from known C^{α} coordinates. The first problem addressed concerns the positioning of the atoms of the first peptide. After this first peptide is positioned the method to reconstruct atomic coordinates

of the backbone and β-carbons involves the use of data bases which contain ranges of protein bond lengths, bond angles and dihedral angles and a Monte Carlo method of selecting values from predetermined ranges along with criteria for their acceptance. We compare the results of the MCPB method with methods previously reported.

3.2. Computational method

Modelling and computations were performed on an IBM RS/6000. The programs *mtranscoil*, *crystal*, *calcsup*, *calcrms* and *range* were written in FORTRAN 77. Statistical analysis was carried out using the program *SAS*[10] running on a VAX750/VMS system. The program *pbdmmod* was used to convert PDB structure files to Macromodel format.[11] Bond lengths, bond angles and dihedral angles were determined with the program *mtranscoil.* A target structure was generated by the program *crystal* and the detail of this program and method is detailed below. The program *range* was used to determine the boundaries of v_6 and v_{13} in the PDB. Energy minimisation was carried out with *Macromodel/Batchmin2K* when the protein contained less than 2000 backbone atoms (i.e. 400 residues) or *Batchmin5K* for larger proteins with up to 5000 atoms e.g. 1TIM, 1PSG, 1CSE and 3RP2. The r.m.s.d.[12] between a model and the crystal structure was calculated using the program *calcsup* which superimposes two structures by the method developed by Kabsch.[13] The program was also used for analysis for each type of amino acid and of the different motif regions.[14] The motif of an amino acid in the analysis was as defined in the PDB.

Two methods are applied to generate the atomic coordinates of the backbone from C^{α} coordinates.

3.2.1. Generation of the first peptide

Our study has focused on the proteins listed in Table 2.1 which have been selected because of their diversity and the accuracy with which the crystal structure data is known. The first problem to address in building the protein concerns the positioning of the atoms of the first peptide and is important since it is from these atoms that further construction occurs. We have chosen to build the protein from the N-amino acid end. The program

crystal chooses which of the terminal amino acid configurations is appropriate from criteria defined later. We first describe building the protein with the first amino acid in a right handed α-helical conformation.

The carbonyl carbon of the first amino acid from the N-terminus has been defined in the following way. A statistical analysis of the α-helical regions of the fourteen proteins (labelled † in Table 2.1) shows the carbonyl carbon, C'_1, of the first amino acid lies close to a plane defined by C_1^{α}, C_2^{α}, C_5^{α} and therefore this atom in the first amino acid is positioned on this plane. Similar analysis of the crystal structures of fourteen proteins, but irrespective of motif, showed the angle[15] C', C_1^{α}, C_2^{α} to be ca 21.5° and therefore this value was chosen in defining the first peptide carbonyl carbon, C' with respect to C^{α}_1 and C_2^{α}. The C'_1, C_1^{α} bond length of the first amino acid was fixed at the value of 1.530 Å (Table 3.1).[16] Solution of three equations[17] that relate the atomic coordinate of C'_1 to the C'_1, C_1^{α} bond length and the angle C', C_1^{α}, C_2^{α} as 21.5° and define the coordinates of C'_1 on the plane C_1^{α}, C_2^{α}, C_5^{α} give rise to four possible solutions for the C'_1 coordinates. The solution that is accepted is chosen to meet the requirement that the distance between C'_1 and C_5^{α} and the distance between C'_1 and C_2^{α} be less than 6.5 Å and 2.8 Å respectively thus requiring C'_1 to lie between C_1^{α} and C_2^{α}.

After the coordinates of carbonyl carbon C'_1 are defined, the atomic coordinates of N_2 and O_1 are calculated from the carbonyl carbon (C'_1), C_1^{α} and C_2^{α} by using the Scheraga algorithm[18] which transforms the internal coordinates into the Cartesian coordinates. For computing the nitrogen coordinates a C'_1, N_2 bond length and C_1^{α}, C'_1, N_2 angle is required and these are randomly selected from within the ranges 1.35 ± 0.05 Å and 117 ± 5° respectively (see Table 3.1). The dihedral angle N_2, C'_1, C_1^{α}, C_2^{α} is set to zero as for a planar peptide bond. For the generation of O_1 coordinates the algorithm requires the carbonyl C'_1 O_1 bond length which is randomly selected within the range 1.23 ± 0.03 Å and an angle for C_1^{α}, C'_1, O_1 selected within the range 122.5 ± 2.5°. The dihedral angle O_1, C'_1, C_1^{α}, C_2^{α} were chosen at 180° as for a planar peptide bond. After establishing the N_2 and O_1 coordinates, the distance between N_2 and C_2^{α} is computed. If the distance is in the range of 1.45 ± 0.04 Å, consistent with a normal bond length, then the coordinates of N_2 and O_1 are accepted and if outside this range the

coordinates of both these atoms are regenerated until the distance is within the defined range.

To establish the coordinates of N_1 the same algorithm is used. A value for the N_1, C_1^{α} bond length is randomly selected from within the range 1.45 ± 0.04 Å and a N_1, C_1^{α}, C'_1 bond angle is similarly selected from within the range of 107 ± 5°. A value for the N_1, C_1^{α}, C'_1, N_2 (ψ_1) dihedral angle is selected from the range -40 ± 20° typical for right handed α-helical proteins.[19] The C^{β} is defined by the same algorithm requiring the distance C_1^{β}, C_1^{α}, randomly selected from the range 1.54 ± 0.04 Å and an angle for C_1^{β}, C_1^{α}, C'_1 similarly selected from the range is 107 ± 5°. The dihedral angle C_1^{β}, C_1^{α}, C'_1, N_2 is ψ_1 + 120° where ψ_1 is chosen from the range -40 ± 20°.

The same method is used for generating the first amino acid in a β-sheet, a β turn or a random coil motif as used in an α-helix configuration except that the assumption is made in all these motifs that the carbonyl carbon C'_1 lies on the C_1^{α}, C_2^{α}, C_3^{α} plane rather than on the C_1^{α}, C_2^{α}, C_5^{α} plane. An analysis of the non helical regions of the fourteen proteins showed this to be close to reality.[20] To establish the coordinates of N_1 and C_1^{β} for these motif regions ψ_1 is chosen from the range, 150 ± 30° for the β-sheet motif. The β-turn and random coil motifs are not differentiated and ψ_1 is chosen from the range 80 ± 50°.[19]

After this first peptide is positioned the method to reconstruct atomic coordinates of the backbone and β-carbons involves the use of data bases which contain ranges of protein bond lengths, bond and dihedral angles and a Monte Carlo method of selecting values from predetermined ranges in values along with criteria for their acceptance.

3.2.2. Generation of the backbone atoms (Method 1)

3.2.2.1. Predetermined ranges

The predetermined ranges of the variables used above for building the first peptide and below from building from the second to the last amino acid of the chain were established from an analysis of the X-ray structures of fourteen proteins in which the total number of C^{α} atoms is 1282 (1CRN, 1CTF, 1PPT, 1RDG, 1ROP, 1UTG, 2CCY, 2OVO, 2ZTA, 3FXN, 3CLA, 3WRP, 4PTI, 4TNC) (labelled with a † in

Table 2.1). The proteins were chosen[21] because the X-ray crystal structure coordinates[22] were known with an accuracy of better than 2.5 Å. Analysis of the fourteen proteins showed the average bond lengths and angles and standard deviations as in Table 3.1 similar to values obtained in other studies.[23]

Table 3.1 The bond length and bond angles used in this work

Bonding atoms	Bond length (Å)	Bonding atoms	Bond angle (°)
C^{α}- C^{β}	1.540 ± 0.04	N - C^{α} - C^{β}	111.0 ± 3.5
C^{α}- C'	1.530 ± 0.04	N - C^{α} - C'	110.0 ± 3.0
C' - O	1.225 ± 0.03	C^{α} - C' - O	122.5 ± 2.5
C' - N	1.340 ± 0.04	C^{α} - C' - N	112.0 ± 5.0
N - C^{α}	1.450 ± 0.04		

3.2.2.2 The definition of ranges of torsion angles

In chapter two we reported the NLRDT method, based on non-linear regression analysis of a protein data base which together with values of the distance $C_{i-1}{}^{\alpha}$, $C_{i+2}{}^{\alpha}$ and torsional angle $C_{i-1}{}^{\alpha}$, $C_i{}^{\alpha}$, $C_{i+1}{}^{\alpha}$, $C_{i+2}{}^{\alpha}$ of the ith amino acid, to assign a specific value of dihedral angle to each amino acid of the protein. The non linear regression equations are surprisingly reliable for predicting the absolute value of the dihedral angles of the protein backbone from the C^{α} coordinates, affording R^2 values of 0.63 for $\cos\phi$ and R^2 values of 0.87 for $\cos\psi$. The sign of the dihedral angles ϕ_i and ψ_i of the ith amino acid is assigned from a comparison of the distance $C_{i-1}{}^{\alpha}$, $C_{i+2}{}^{\alpha}$ and the torsional angle $C_{i-1}{}^{\alpha}$, $C_i{}^{\alpha}$, $C_{i+1}{}^{\alpha}$, $C_{i+2}{}^{\alpha}$ with predetermined ranges of these values established to best correlate with sign for the protein data base used.

These values of the dihedral angles for each amino acid of the protein obtained in this way cannot however in themselves be used to define coordinates of the backbone atoms since they result in unacceptable bond lengths and angles of the backbone atoms in relation to the defined C^{α} coordinates. However if it is assumed, as our study has

showed, that the dihedral angles of the protein have a high probability of occurring within a window of ± 30° of the computed value, then angles from these ranges can be selected for each residue of the protein by Monte Carlo methods. In accepting the selected values, criteria must be met which allow for the positioning the backbone atoms with appropriate bond lengths and angles.

3.2.2.3 The second and subsequent amino acids

The second amino acid from the N-amino end is built to the first amino acid (whether as an α-helical configuration or a non-helical configuration) in the following way. The coordinates of the nitrogen and C^{α} of the second amino acid are already defined. The dihedral angle ϕ_2 will therefore define the positions of C^{β} and the carbonyl carbon of the second amino acid along with appropriate bond lengths and bond angle. In the same way the dihedral angle ψ_2 and appropriate bond lengths and angles will define the positions of the nitrogen of the third amino acid and the carbonyl oxygen of the second amino acid. A value for each of the dihedral angles ϕ and ψ for the amino acid is selected by a random number method from the ranges defined by a ± 30° window of the computed value using in the first instance of the NLRDT method. Appropriate bond lengths and angles are selected by a random number method from the values defined in Table 3.1. If the resulting coordinates meet the following criteria they are accepted.

3.2.2.4. Criteria of acceptance of coordinates

After the coordinates of a peptide unit are generated, and before the next peptide unit is built, the question arises as to whether the coordinates be accepted or rejected. Two criteria for acceptance have been found effective and depend on values of the distance d_i and torsional angle ω_i (Figure 3.1):

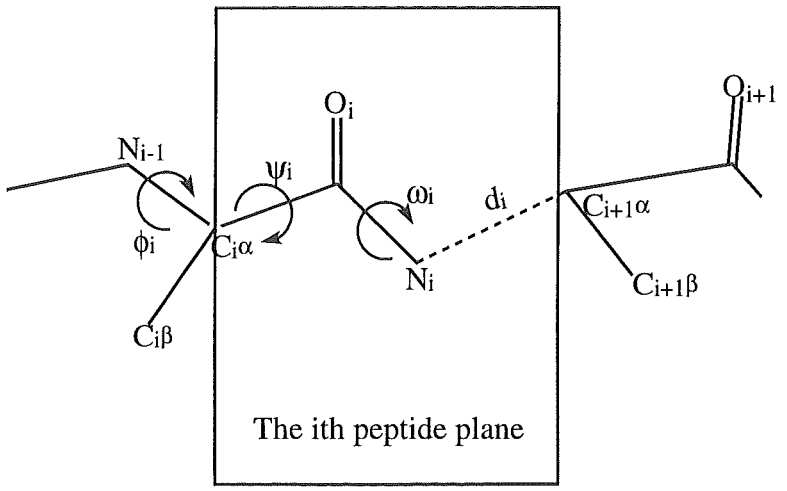


Figure 3.1 Criteria of acceptance of atomic coordinates

(i) the bond length of the ith peptide unit, d_i, between N_i in the ith peptide and $C_{i+1}{}^{\alpha}$ of next peptide must be in range 1.40 Å $\leq d_i \leq$ 1.50 Å and (ii) the torsion angle, ω_i for a *trans*-peptide $|\omega_i| \geq 170^{\circ}$ and for a *cis*-peptide $|\omega_i| \leq 10^{\circ}$. The difference in distance between $C_i{}^{\alpha}$ and $C_{i+1}{}^{\alpha}$ with the absolute value of ω allow the *cis*- and *trans*- peptide configurations to be recognised. Analysis of the fourteen proteins showed the distance between $C_i{}^{\alpha}$ and $C_{i+1}{}^{\alpha}$ for a *trans*-peptide is in the range 3.80 ± 0.35 Å and for a *cis*-peptide 2.8 ± 0.65 Å. If the values of d_i and ω_i are within the criteria, the coordinates are accepted and the atomic coordinates of next peptide generated. If the values of d_i or ω_i are outside these specified ranges then the coordinates of atoms in this peptide unit are rejected and the random seed generator is used to select alternate values and the process repeated. If suitable values of d_i or ω_i are not determined within 500 cycles the criteria conditions of d_i or ω_i are relaxed by 0.01Å and 1° respectively until such time as coordinates are acceptable. This treatment, ensures that the process of generating atomic coordinates from the second amino acid residue to the last one can continue. In our studies a suitable set of coordinates that meet the criteria has always been obtained without resorting to relaxing the criteria. Using this method to regenerate the backbone atoms from the C^{α} coordinates a sample of thirteen proteins was examined (Table 3.5) The average r.m.s.d of the backbone was calculated as 0.59 Å. This average error resulting from selecting dihedral angles from within the ± 30° of the angle calculated

from the NLRDT method was considered outside of an acceptable error and an alternative data base from which dihedral angles could be selected was developed.

3.2.3. Generation of the backbone atoms (method 2)

3.2.3.1. The definition of range of dihedral angle

The second method of developing a data base for dihedral angles was established from a statistical analysis of the relationship between dihedral angles of the backbone and C^{α} coordinates for a sample of fourteen proteins using SAS. The coefficients of the correlation matrix between the dihedral angles of the backbone and v_1-v_{15} parameters of the C^{α} coordinates showed the $C_{i-1}{}^{\alpha}$ and $C_{i+2}{}^{\alpha}$ distance (v_6) and the torsional angles $C_{i-1}{}^{\alpha}$, $C_i{}^{\alpha}$, $C_{i+1}{}^{\alpha}$, $C_{i+2}{}^{\alpha}$ (v_{13}), to be important variables in assignment of secondary structure motif (see table 2.3 in chapter two). For the selected proteins the distances between $C_{i-1}{}^{\alpha}$ and $C_{i+2}{}^{\alpha}$ (v_6) range from 4.0 to 13.0 Å and the torsional angles $C_{i-1}{}^{\alpha}$, $C_i{}^{\alpha}$, $C_{i+1}{}^{\alpha}$, $C_{i+2}{}^{\alpha}$ (v_{13}) range through 360°. The dihedral angles are recorded from 180° to -180° where positive values represent a clockwise twist in the direction of the protein from the N-terminal to the C-terminal.

We were interested to see if there is a relationship between values of v_6 and v_{13} and ϕ and ψ that would be better than the NLRDT regression method in providing a data base from which dihedral angles could be selected in modelled the backbone. For the fourteen protein selection we have investigated the distribution of ϕ_i and ψ_i for the ith amino acid with v_6 and v_{13} for that residue (see Figure 3.2). We first divided the values of v_6 and v_{13} into four arbitrary ranges: v_6 with boundaries (1) 4.0 Å to 6.0 Å, (2) 6.0 Å to 8.0 Å, (3) 8.0 Å to 9.5 Å, (4) 9.5 Å to 13.0 Å respectively and v_{13} into quadrants (1) 180° to 90°, (2) 90° to 0°, (3) 0° to -90°, (4) -90° to - 180° respectively. For example, for the ith amino acid residue, if the distance, v_6 (between $C_{i-1}{}^{\alpha}$ to $C_{i+2}{}^{\alpha}$), is in the range 4.0 Å - 6.0 Å, then v_6 is defined to be in range (1) and if the torsion angle, v_{13}, ($C_{i-1}{}^{\alpha}$, $C_i{}^{\alpha}$, $C_{i+1}{}^{\alpha}$, $C_{i+2}{}^{\alpha}$) is between 90° to 0°, v_{13} is defined to be in range (2). The $C_i{}^{\alpha}$ atom is then recorded as in range (1) for v_6 and range (2) of v_{13}. For the sake of simplification this is recorded as the REGIONs 12 - the first number of a REGION representing a range

of v_6 and the second the v_{13} value. There are sixteen such REGIONs from 11 through 44.

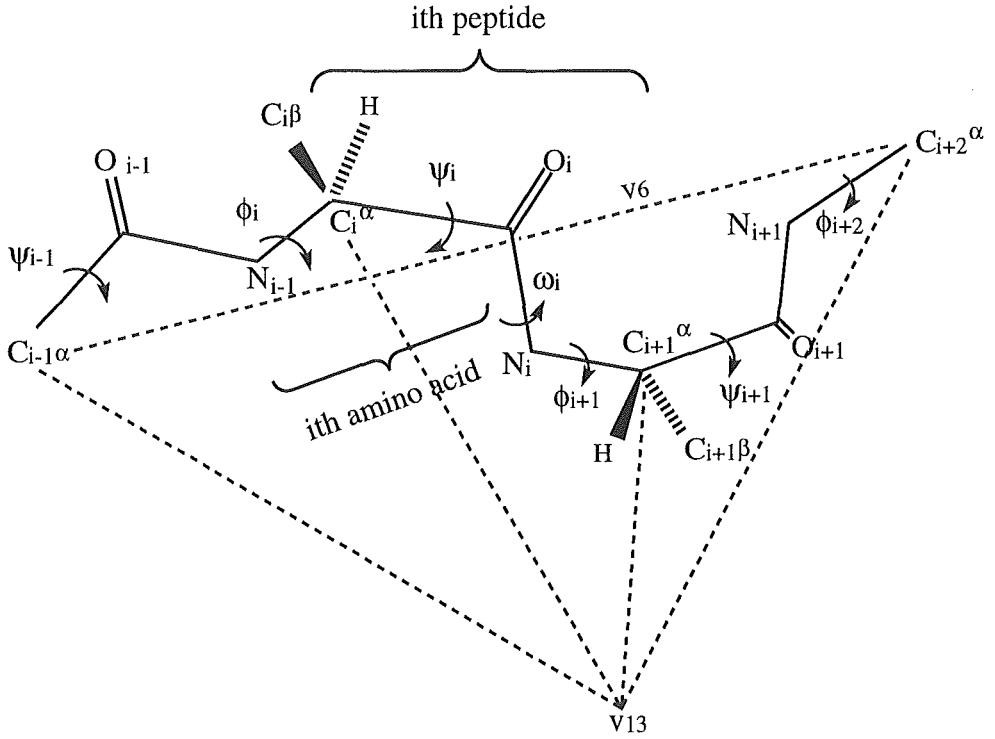


Figure 3.2 The demonstration of v_6 and v_{13} related to ϕ and ψ

The mean values (ϕ and ψ) and standard deviations of the dihedral angles of amino acids which have values of v_6 and v_{13} corresponding to the ranges of v_6 and v_{13} defined in the REGION 11 to 44 above were determined for the fourteen proteins in the data base. The arbitrary boundaries of v_6 and v_{13} defined above for REGIONs 11 to 44 were adjusted so as to give a minimum overall average standard deviation in the values of ϕ and ψ. It was hoped that by this procedure an optimised set of boundaries of the values of v_6 and v_{13} defining the REGIONs 11 to 44 would allow a dihedral angle for each amino acid of the protein to be selected by Monte Carlo methods and maximise the change of building protein model with a low r.m.s.d. The arbitrary boundaries of v_6 and v_{13} which define the REGIONs 11 to 44 were adjusted by a random procedure. The boundary for the ranges were randomly varied by $\pm$ 1 Å for v_6 and $\pm$ 40° for v_{13}. For example for v_6 a range of 4 - 6 Å, 6 - 8 Å, etc. can be changed to become 4 - 5 Å, 5 - 8 Å or 4 - 7 Å, 7 - 8

Å etc. More than 1000 random selections of boundaries were examined. The overall average of the standard deviations of the dihedral angles for the amino acids which have values of v_6 and v_{13} corresponding to the new boundaries were determined for the fourteen proteins in the data base using the program *range*. The boundaries which gave rise to the lowest overall standard deviation is defined below in Table 3.3. No dihedral angles were observed for the proteins in the REGIONs 11, 14 and 42. Values of ϕ and ψ in these REGIONs as reported in Table 3.3 are taken from analysis of the twenty four protein sample along with literature reports of dihedral angles in β-sheets[19]

Table 3.3. Mean and standard deviations in the values of ϕ and ψ in the REGIONs 11 to 44*

	$-\phi$			$+\phi$			$-\psi$			$+\psi$		
REGION	u	σ	P‡	u	σ	P‡	u	σ	P‡	u	σ	P‡
12	-63.6	7.8	99.6	66.1	32.1	0.4	-40.1	10.3	99.3	91.6	52.6	0.7
13	-62.3	†0.0	33.3	35.6	25.7	66.7	0.00	0.00	0.00	46.1	10.3	100
21	-104.2	13.3	68.2	73.0	10.4	31.8	-10.1	0.00	4.5	26.9	16.1	95.5
22	-82.7	21.9	92.4	101.6	28.0	7.6	-37.7	29.7	77.2	81.6	64.9	22.8
23	-84.6	16.3	91.7	93.2	13.2	8.3	-18.8	14.8	38.9	104.9	70.7	61.1
24	-82.9	17.7	100.0	0.00	0.00	0.0	-17.4	11.9	66.7	4.7	1.58	33.3
31	-98.8	19.9	58.0	76.9	18.5	42.0	-18.3	20.4	12.0	39.2	34.6	88.0
32	-112.3	17.8	57.1	91.5	5.50	42.9	-8.8	2.1	71.4	5.5	1.03	28.6
33	-87.6	29.5	90.0	106.0	32.2	10.0	-66.9	68.5	10.0	154.4	13.8	90.0
34	-87.6	23.9	97.7	168.9	0.00	2.3	-19.5	11.5	22.7	118.8	37.4	77.3
41	-109.3	25.4	80.5	88.0	17.9	19.5	-8.3	7.2	7.3	100.2	49.0	92.7
43	-112.9	32.2	96.9	167.5	0.00	3.1	-168.5	6.7	6.1	156.9	11.9	93.9
44	-111.1	27.3	98.4	51.8	10.3	1.6	-79.1	66.6	2.5	137.7	22.9	97.5
§Av σ		21.1			19.4			22.7			29.8	

* Definitions of REGIONs of v_6 and v_{13} which result in minimum overall average standard deviation in values of ϕ and ψ are; v_6 (1) 4.0 Å - 5.84 Å, (2) 5.84 Å - 8.09 Å,

(3) 8.09 Å - 9.10 Å, (4) 9.10 Å - 13.0 Å . v_{13} (1) 180° - 99.62°, (2) 99.62° - -8.97°, (3) -8.97° - -118.93°, (4) -118.93° - -180°. † Only a single value in the data base for the fourteen proteins. ‡ Probability that a dihedral angle of a particular amino acid from the 14 proteins will occur in this range. § The overall standard deviations for all the regions is (21.1+19.4+22.7+29.8)/4 = 23.25°.

The standard deviations in the ranges of the dihedral angles reported in Table 3.3 vary considerably. The boundaries of the REGIONs 11 to 44 from which ϕ and ψ values are taken were further optimised by repetitive small empirical changes to the centre of the v_6 and v_{13} ranges. This was effected using the program *crystal.*

Table 3.4. Range in negative and positive values of ϕ and ψ for each region

	ϕ		ψ	
REGION	Range 1[a]	Range 2[b]	Range 1[c]	Range 2[d]
11	-45±25	45±25	-5±40	5±40
12	-62±24	35±50	-40±20	90±60
13	-50±25	36±50	47±30	-90±60
14	45±35	-45±35	30±30	-30±30
21	-105±40	75±35	30±35	-10±35
22	-80±40	100±50	-40±50	80±60
23	-85±40	95±30	105±65	-20±30
24	-80±40	80±40	-20±30	10±25
31	-100±40	77±45	40±60	-20±40
32	-110±40	90±25	-10±20	10±40
33	-90±50	105±50	155±25	-70±60
34	-90±50	170±10	120±40	-20±33
41	-110±50	90±35	100±55	-10±20
42	-130±40	125±45	155±25	150±30
43	-110±40	150±35	155±30	-170±10

44	-110±50	50±30	140±50	-80±60

[a] more probable sign of ϕ; [b] less probable sign of ϕ ; [c] more probable sign of ψ; [d] less probable sign of ψ .

The total number of variations examined was greater than 4000. The ranges which allowed for prediction of a backbone structure of the twenty four proteins by the MCPB method with a minimum r.m.s.d's are shown in Table 3.4. Values of ϕ and ψ for an amino acid residue reflect the twist of the protein and can be both positive, both negative or one positive and the other negative. The Table 3.4 shows the preference for a positive or negative values of the dihedral angle.

3.2.3.2. Building the backbone from the first amino acid

In building the coordinates of the backbone values of v_6 and v_{13} for each amino acid residue are established from the C^{α} coordinates with the subroutine *cavara* in *crystal.* The REGION 11 to 44 for each C_i^{α} is then established on the basis of these values in Table 3.4. This also includes the first amino acid, the coordinates of which are then established using the method described above.

The second amino acid from the N-amino end is built to the first amino acid (whether as an α-helical configuration or a non-helical configuration) in the following way. As before the coordinates of the nitrogen and the C^{α} of the second amino acid are already defined. The dihedral angle ϕ_2 will therefore define the positions of C^{β} and the carbonyl carbon of the second amino acid along with appropriate bond lengths and bond angles. In the same way the dihedral angle ψ_2 and appropriate bond lengths and angles will define the positions of the nitrogen of the third amino acid and the carbonyl oxygen of the second amino acid. A value for each of the dihedral angles ϕ and ψ for the amino acid is randomly selected according to the assigned REGION taking into account the more probable sign combination of the dihedral angles (Table 3.4). Appropriate bond lengths and angles are selected by a random number method from the values defined in Table 3.1.

If the resulting coordinates meet the criteria defined earlier in the section 3.2.2.4 they are accepted.

In building a residue, i.e. the second or the ith residue, it is possible to choose +ve or -ve values for ϕ and ψ. Of the four combinations (++, --, +-, -+) the probability is not the same. An analysis of the probability of a positive or negative value of ϕ and ψ for the fourteen chosen proteins is recorded in Table 3.4. In reconstructing each amino acid of the protein the choice of sign of ϕ and ψ can be (i) the more probable sign of ϕ and ψ or (ii) the more probable sign of ϕ and less probable of ψ (iii) less probable sign of ϕ and more probable of ψ (iv) less probable sign of ϕ and less probable of ψ. The program *crystal* is written such that, in the building of the protein, combination (i) is chosen first and if the criteria for accepting the coordinates so produced is not met within 500 cycles then combination (ii) is chosen and so on. If suitable values of d_i or ω_i are not determined within 1000 cycles the next most probable combination of sign of ϕ and ψ are used to generate the coordinates and so on. If suitable values are not found within 1500 cycles then the final combination is used and if suitable values are not found within cycles 2000 then the program returns to the most probable sign combination but the criteria are relaxed by 0.01 Å for d_i or 1° for ω_i, e.g. 1.39 Å $\leq d_i \leq$ 1.50 Å and $|\omega_i| \geq 169°$ or $|\omega_i| \leq 11°$ until 1.35 Å $\geq d_i$ or $d_i \geq$ 1.55 Å and $|\omega_i| \leq 140°$ or $|\omega_i| \geq 40°$ for *trans* and *cis*- peptide are met. After a set of coordinates are accepted the acceptance ranges for d_i and ω_i are reset to their initial default values and most probable sign combination before generating the next peptide unit. The process is repeated for each amino acid residue of the backbone of the protein.

Since the C^{α} parameters can not be defined at the carbonyl terminus of the last residue the terminal N atom is capped with a methyl group. The procedure described is encapsulated in the program *crystal* and the method we describe as Monte Carlo Protein Building (MCPB).

3.2.4. Mean coordinates of the backbone model

Levitt was the first to average coordinates from several model structures of a protein.[24] We have similarly shown that the r.m.s.d of an average or mean structure of

several model structures with the X-ray structure was smaller than for any individual model. As a result we have examined this methodology in more detail. The bond lengths and bond angles in the averaged model structure are checked for distortions and to see if they are in the defined ranges as in Table 3.1. If they are outside this ranges *crystal* is programmed to recognise the structure as distorted. If a structure was distorted energy minimisation could be used to bring the bond lengths and angles to sensible values. To date we have not found any averaged structure of protein backbone to be distorted.

An empirical study was made to determine if there was an optimum number of independent models that should be averaged to give a mean structure with the lowest energy r.m.s.d. It was found that little benefit is gained in averaging more than ten structures.

3.2.5. Energy minimisation of the backbone model

An average structure (of ten structures) without hydrogen atoms for each of the twenty-four proteins was minimised with *Macromodel/Batchmin* using the OPLS/AMBER (non-hydrogen) force field[25] and GB/SA water model.[26] The program uses default values of the cutoff distances of non-bonded interactions and electrostatic interactions. The C^{α} atoms of the model structure are restrained to their initial positions as defined by the X-ray coordinates.[27]

We examined the effect of removing the constraints on the C^{α} atomic positions during minimisation for the mean structures. When the constraints on the C^{α} positions are released the coordinates of the C^{α} positions move and the r.m.s.d necessarily increases, by inclusion of the now variable C^{α} positions, to ca 1.1Å. The results show that after minimisation without constraint on the C^{α} positions the r.m.s.d. is substantially increased. For the subsequent studies it was considered pertinent to effect minimisation with the C^{α} constraints in place.

3.3. Results and discussion

3.3.1. The r.m.s.d. of the backbone model

The MCPB method has been applied to the twenty four proteins (Table 2.1) containing 3,563 residues and ranging is size from 36 to 494 residues. An advantage of the method is that it is capable of generating the structure of proteins that contain sub-units or peptide chains that are not linked. For example 1CSE, 1PSG, 1TIM, 2CCY, 2ZTA and 3RP2 each contain two chains or two sub units that are not linked while 1AMT contains three non linked chains and 4INS contains four non linked chains. The protein 2ZTA is a parallel coiled-coil two chain protein and 1ROP is an anti-parallel coiled-coil protein.

Table 3.5 The average atomic deviation and r.m.s.d. of the backbone models averaged

Protein	N[a]	av-dev[b]	r.m.s.d.[c]	av-dev.[d]	r.m.s.d.[e]	r.m.s.d.[f]
1AMT	60	0.26	0.41	0.18	0.36	-
1CRN	46	0.39	0.63	0.23	0.47	0.54
1CSE	337	0.39	0.59	0.27	0.55	-
1CTF	68	0.38	0.59	0.24	0.48	0.73
1PCY	99	0.44	0.64	0.26	0.53	-
1PPT	36	0.40	0.59	0.18	0.39	-
1PSG	365	0.39	0.59	0.29	0.55	-
1RDG	52	0.39	0.57	0.23	0.43	-
1ROP	56	0.23	0.36	0.14	0.30	0.31
1TIM	494	0.37	0.56	0.29	0.52	-
1UTG	70	0.27	0.38	0.14	0.24	0.47
2CCY	254	0.29	0.48	0.19	0.41	0.58
2LYM	129	0.42	0.67	0.31	0.61	-
2LYZ	129	0.41	0.67	0.32	0.61	-
2MHR	117	0.33	0.57	0.19	0.48	-
2OVO	56	0.36	0.50	0.19	0.33	0.86
2ZTA	62	0.21	0.28	0.14	0.22	0.23
3FXN	138	0.34	0.52	0.23	0.43	0.79

3CLA	125	0.32	0.47	0.18	0.33	0.57
3RP2	448	0.39	0.57	0.25	0.49	-
3WRP	101	0.23	0.37	0.15	0.29	0.43
4INS	102	0.39	0.64	0.23	0.50	0.82
4PTI	58	0.41	0.60	0.29	0.57	0.87
4TNC	160	0.28	0.40	0.17	0.28	0.55
overall		0.35	0.53	0.22	0.43	0.59

[a] The number of residues in the protein; [b] average deviation (Å) before energy minimisation; [c] r.m.s.d. (Å) before energy minimisation of the model structure; [d] average deviation (Å) after energy minimisation; [e] r.m.s.d. (Å) after energy minimisation; [f] r.m.s.d. (Å) after energy minimisation of the model structures which were obtained by using the method 1.

The average atomic deviation and the r.m.s.d. of the backbone and C^{β} atoms of the mean modelled structures and the X-ray structures for the twenty-four proteins are listed in the Table 3.5. For the backbone atoms the average atomic deviation is 0.35 Å and r.m.s.d. 0.53 Å and these values reduce to 0.22 Å and 0.43 Å respectively after energy minimisation. Energy minimisation with the C^{α} atoms fixed reduces the r.m.s.d by ca. 0.1 Å. The proteins with the higher content of helical structure, e.g. 1AMT, 3WRP and 4TNC and the coiled-coil proteins, 2ZTA and 1ROP give the lowest r.m.s.d. values. The backbone atoms, nitrogen, carbonyl carbon, oxygen including C^{β} have r.m.s.d. of 0.29 Å, 0.26 Å, 0.77 Å, and 0.36 Å, respectively. Oxygen shows a greater variation than C^{β}.

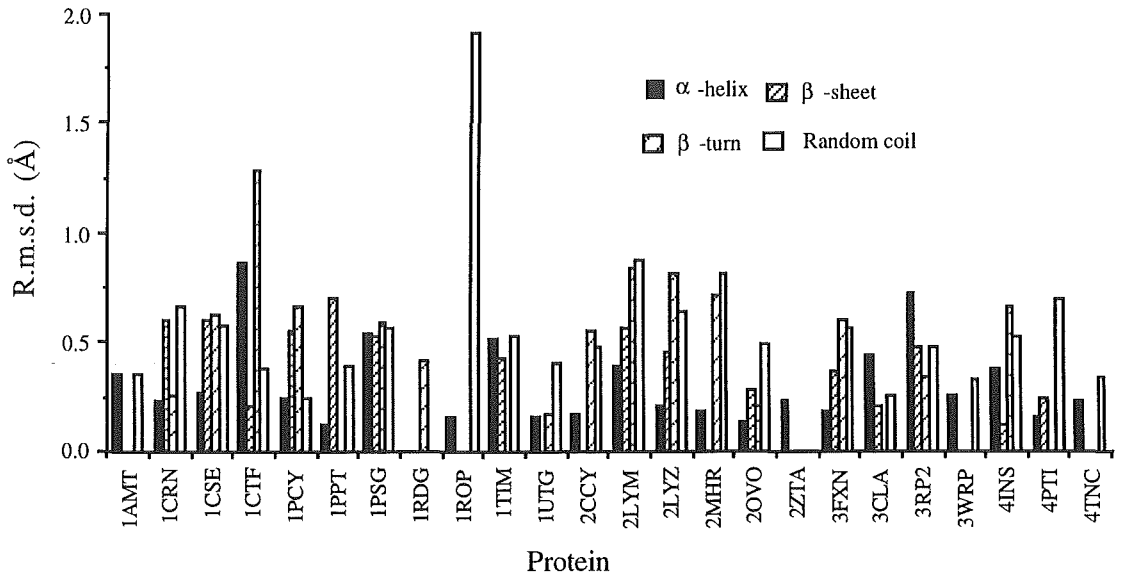


Figure 3.3 The r.m.s.d for amino acids in each motif region in the twenty four proteins.

The mean model structures of each of the twenty-four proteins were examined after energy minimisation by motif and the r.m.s.d for the amino acids in each motif region calculated and shown in Figure 3.3. In general the r.m.s.d of the secondary structure is lowered by energy minimisation with the C^{α} coordinates fixed. Regions of α-helix, β-sheet and β-turn are as defined in the PDB file and the remainder of the structure is assumed to be random coil.

The average r.m.s.d. of α-helical, β-sheet, β-turn region and random coil regions are 0.31 Å, 0.42 Å, 0.58 Å and 0.57 Å respectively. The α-helical region gave the lowest deviation. The r.m.s.d of β-sheet regions is lower than the β-turn and random coil regions but higher than the average r.m.s.d.

The r.m.s.d of each of the twenty amino acids in the averaged models of the twenty-four proteins are plotted in Figure 3.4. The number of amino acids included in the analysis is 3536. (There are twenty-seven undefined amino acid residues in the protein 1AMT that have no definition in it's PDB file.) The overall average r.m.s.d after energy minimisation of the averaged structures for all the amino acids was 0.47 Å. Glu (E) showed the minimum r.m.s.d. (0.37 Å) and Glycine (G) the maximum (0.66 Å). The total number of proline residues in the twenty-four proteins is 151 (141 *trans*- and 10 *cis*-

). Proline showed an r.m.s.d of 0.53 Å (*trans* 0.53 Å and *cis* 0.61 Å) after energy minimisation (see Figure 3.4).

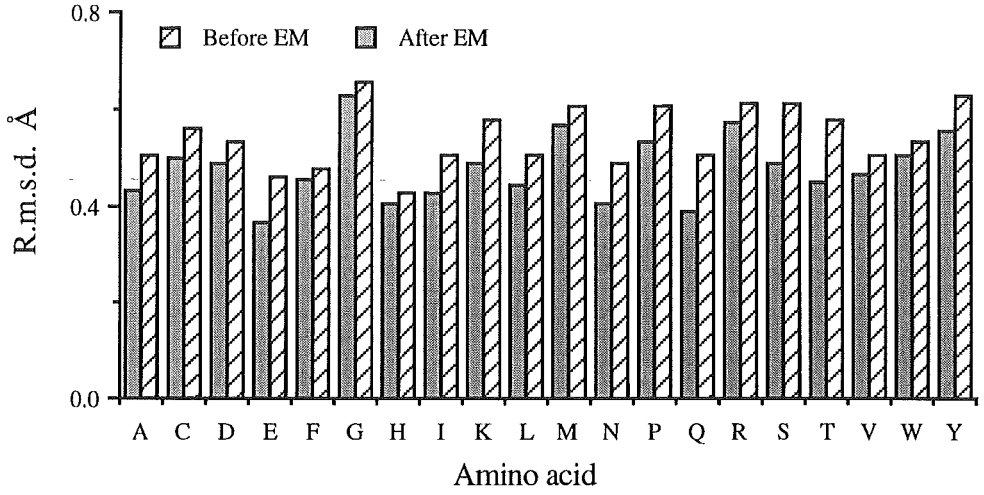


Figure 3.4. The r.m.s.d of the amino acids in the twenty four proteins.

3.3.2. The dihedral angle of the backbone model

The percentage of the dihedral angles after energy minimisation of the averaged model structures that fall within in a ± 30° window of the X-ray structure for the twenty-four proteins is given in figure 3.5.

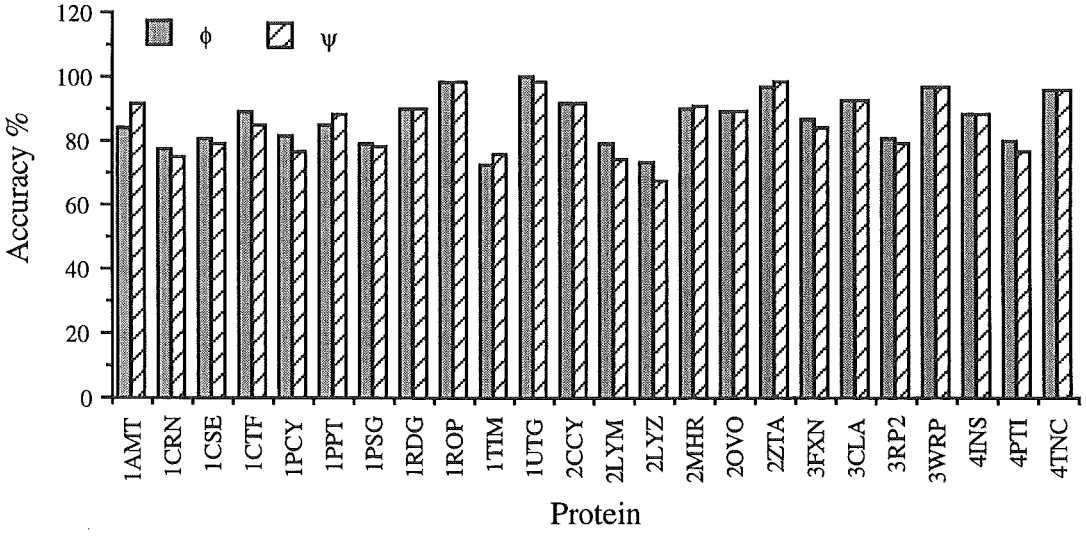


Figure 3.5. Percentage of dihedral angles of the amino acids within a ± 30° window of the value in the X-ray structures for 24 proteins.

Energy minimisation of the structure increases the percentage of values of ϕ and ψ within a ± 30° window. The overall percentage of values of ϕ and ψ within a ± 30° window for the twenty-four proteins is 87% and 86% respectively (Figure 3.5). Proteins with a high helical content, e.g. 2ZTA, 1ROP, 1UTG, 3WRP and 4TNC give a high percentage > 90% of dihedral angles within the ± 30° window. For the proteins with a high content of β-turn and random coil segments, e.g. 1CRN, 1PCY, 1PSG, 1TIM, 2LYM, 2LYZ, 4PTI, 4INS and 3RP2, the accuracy of prediction within the window is below 85%.

The percentage of dihedral angles for each amino acid within a ± 30° window of the value from the X-ray structure in shown in Figure 3.6 for the each of the amino acids in twenty-four proteins. The overall percentages of dihedral angles ϕ and ψ for the twenty amino acids within a ± 30° window of the values found from the X-ray structure are 84% and 83% respectively. For most amino acids 80% of the values of ϕ and ψ fall within a ± 30° window. The greatest deviation is observed for the dihedral angles of glycine, where there are no side-chain constraints, where 75% of the values of ϕ and 72% of the values of ψ fall within the ± 30° window. For proline, 83% of the values of ϕ and 83% of the values of ψ respectively fall within the ± 30° window. The value for *cis*- and *trans*-proline are comparable. Proline therefore parallels the other amino acids. In all cases energy minimisation has the effect of increasing the percentage of values of ϕ, ψ that fall within the ± 30° window by ca 10%.

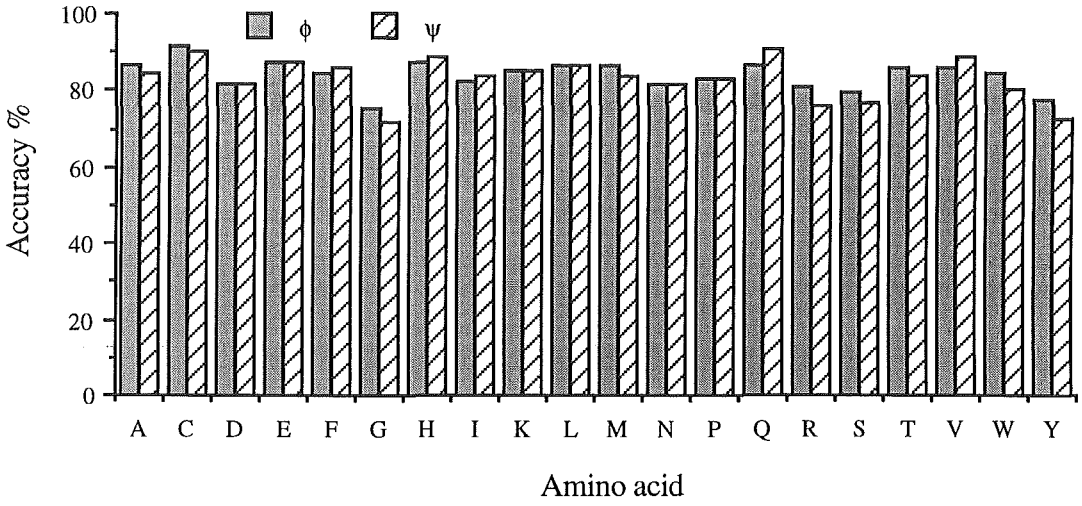


Figure 3.6. Percentage of dihedral angles for each amino acid within a ± 30° window of the value in the X ray structure for 20 amino acids.

3.3.3. Comparison of the results with previous methods

In comparing the ability of the MCPB method to model protein coordinates with these other methods three criteria are considered: the r.m.s.d. of the backbone and C^{β} atoms with the X-ray structure, the information required for the data base to build the backbone and the computational requirements.

The method pioneered by Jones and Thirup[6] uses a data base or dictionary of structural templates from known structures. Reid and Thornton[3] applied this method to the generation of the main chain atomic coordinates of 3FXN and obtained a r.m.s.d for the main chain 0.51Å. Claessens *et al*[28] was the first to use an automatic search segment data base for building the backbone atoms of three proteins, triose phosphate isomerase, citrate synthase and carboxyl peptidase and obtained an average r.m.s.d of 0.61 Å, somewhat poorer than the average r.m.s.d (0.43 Å) obtained in the present studies.

A method called 'segment match modelling' (SMM) which requires the amino acid sequence and a data base of X-ray structures has been developed by Levitt.[24] The method also positions C^{β} atoms. An average r.m.s.d of the main chain atoms excluding C^{β} atoms

of 0.42 Å for the modelling eight proteins was obtained. The method gave a satisfactory result for 1CTF and 4PTI, but the results for 1CRN and 3FXN are not as good.

Recently, Mandal has developed PROGEN[9] that uses an optimal geometry parameter database established from a statistical analysis employing twenty-three different intra-peptide and inter-peptide geometric parameters relating to C^{α} coordinates for nineteen proteins for the positioning of atoms for each amino acid. The r.m.s.d for the backbone atoms and C^{β} ranged from 0.29 Å - 0.76 Å (average value of 0.53 Å) and with average atomic deviations between 0.14 Å and 0.44 Å (average 0.28 Å).

Correa[4] used molecular dynamics to build protein structures from the C^{α} positions in a method that does not use any data base of known protein structural information. The r.m.s.d. of backbone atoms for three proteins (α-lytic protease, troponin C and flavodoxin) using this method are 0.19 Å, 0.41Å and 0.49 Å respectively. The method requires considerable computer time.

Recently, Rey and Skolnick developed a purely analytical method[7] for generating a protein backbone from C^{α} coordinates and used the method to built the backbone structures of six proteins. The average r.m.s.d. of the backbone atoms was 0.70 Å before energy minimisation, and 0.49 Å after energy minimisation. A comparison of the results from different methods is shown in Table 3.6.

The amino acid sequence of a protein is not required in the generation of the model of backbone atoms including C^{β} atoms in our MCPB method however the sequence is required for positioning of the side-chains. This has its value. For example, the protein 1AMT has 61 amino acid residues and 27 of them have no definition in the PDB file. The MCPB method can generate the backbone atomic coordinates, including C^{β} atoms, from the C^{α} coordinates and give a model with a reasonable r.m.s.d. This method at this stage does require a full set of C^{α} coordinates. The method also has an advantage of being computer inexpensive. For the generation of an average model using the program *crystal* approximately 0.5s of computer time is required per residue. Energy minimisation requires approximately 9s of computer time per residue.

Table 3.6 The r.m.s.d. after energy minimisation for eight different modelling methods

	R.m.s.d (Å)							
Protein	Method 1	2	3	4	5	6	7	8
3FXN	0.43	0.51	0.48	0.49	0.44	0.49	0.46	0.51
1CTF	0.48	-	-	-	0.29	-	-	-
1CRN	0.47	-	-	-	0.56	-	-	0.43
1TIM	0.52	-	0.59	-	-	-	0.53	-
4TNC	0.28	-	0.41	-	-	-	0.36	-
4PTI	0.57	-	-	0.43	0.51	-	0.45	-

Method 1. MCPB in this work. 2. Reid and Thornton (1989). 3. Holm and Sander (1991). 4. Rey and Skolnick (1992). 5. Levitt (1992). 6. Correa (1990). 7. Mandal and Linthicum (1993). 8. Mathiowetz and Goddard (1995).

3.4. Conclusion

In this chapter, we have presented a MCPB method that allows for the construction of a protein backbone from C^{α} coordinates. The method requires the C^{α} coordinates to be known but the sequence is not required. The method gives the backbone structures whose coordinates deviate from the X-ray coordinates by an average of 0.52 Å before energy minimisation, and 0.43 Å after energy minimisation with the OPLS/Amber force field and GB/SA solvent method comparing favourably with other methods. The computational time to generate the backbone coordinates is cost effective. The method is not demanding in computer time even with inclusion of the energy minimisation. The method is accurate, efficient and robust for the modelling of protein backbones. The modelling technique has the potential when integrated into and used in conjunction with traditional X-ray techniques to speed up structure solution.

A simulated annealing procedure is reported in chapter four to predict the side-chain conformation.

References

1 McCammon J. A., Harvey S. C. Dynamics of Proteins, Nucleic Acids Cambridge Unversity Press, (1987) Cambridge, UK.

2 Purisima E. O., Scheraga H. A. (1984) *Biopolymers* **23,** 1207-1224.

3 Reid L. S., Thornton J. M. (1989) *Proteins* **5,** 170-182.

4 Correa P. E. (1990) *Proteins* **7,** 366-377.

5 Holm L., Sander C. (1991) *J. Mol. Biol.* **218,** 183-194.

6 Jones T. A., Thirup S. (1986) *EMBO J.* **5,** 819-822.

7 Rey A., Skolnick J. (1992) *J. Comput. Chem.* **13,** 443-456.

8 Mathiowetz A. M., Goddard W. A. (1995) *Protein Science* **4,** 1217-1232.

9 Mandal C., Linthicum, D. S. (1993) *J. Computer-aided Design* **7,** 199-224.

10 DiIorio, F. C. SAS Application Programing: A Gentle Introduction, Duxbury Press U.S.A. (1991).

11 Macromodel package, Department of Chemistry, Columbia University, New York, NY 10027

12 r.m.s.d. means of root-mean-square deviation.

13 Kabsch W. A. (1976) *Acta Crystallogr.* **A32,** 922-923. Kabsch. W. (1978) *Acta Crystallogr.* **A34,** 827-828.

14 The results of r.m.s.d. calculations were checked using *PCmodel* software. PCMODEL 4.0 version, Serena Software, Box 3076 Bloomington, IN 47402-3076.

15 Leslie A. G. W. (1990) *J. Mol. Biol.* **213,** 167-186.

16 Nemethy G., Pottle M. S., Scheraga H. A. (1983) *J. Phys. Chem.* **87,** 1883-1887.

17 The three equations are expressed as :

(1) $p_1^2 + q_1^2 + r_1^2 = d_1^2$

(2) $p_1p^2 + q_1q^2 + r_1r_2 = (\cos\alpha_1)[(p_1^2+ q_1^2+r_1^2)(p_2^2+q_2^2+r_2^2)]^{-1/2}$

(3) $Ap_1 + Bq_1 + Cr_1 = (\sin\alpha_2)[(A^2+B^2+C^2)(p_1^2+ q_1^2+r_2^2)]^{-1/2}$

Where the p_i, q_i and r_i ($i = 1, 2$) are vectors of the atomic positions; the distance d_1 is the bond length, C', C_1^{α}; α_1 the angle C', C_1^{α}, C_2^{α} with a value of 21.5°; α_2 the angle between the line of C', C_1^{α} and the plane formed by C_1^{α}, C_2^{α}, C_5^{α} has a value 0.0°; A, B and C are the parameters for the equation of the plane formed by C_1^{α}, C_2^{α}, C_5^{α}. The vectors and the plane parameters A, B and C are obtained from ;

$p_i = x - x_i,$

$q_i = y - y_i,$

$r_i = z - z_i,$

$A = (y_2-y_1)(z_3-z_1) - (y_3-y_1)(z_2-z_1),$

$B = (x_3-x_1)(z_3-z_1) - (x_2-x_1)(z_3-z_1),$

$C = (x_2-x_1)(y_3-y_1) - (x_3-x_1)(y_2-y_1),$

x, y, z, are the coordinates of the unknown atom. x_i, y_i, z_i are the coordinates of the known atoms.

18 Dudek M. J., Scheraga H. A. (1990) *J. Comp. Chem.* **11,** 121-151.

19 Levitt M., Greer J. (1977) *J. Mol. Biol.* **114,** 181-293. Spera S. Bax A. (1991) J. Am. Chem. Soc. **113**, 5490-5492.

20 An analysis of non-helical structure motifs in fourteen proteins show the C'_1, O_1 and N_2 plane makes a dihedral angle with the C_1^{α}, C_2^{α}, C_3^{α} plane of 20° ± 50° and therefore the assumption has limitations.

21 Bernstein F. C., Koetzle T. F., Williams E. J. B., Meyer Jr. E. F., Brice M. D., Rodgers J. R., Kennard O., Shimanouchi T., Tasumi M. (1977) *J. Mol. Biol.* **112,** 535-542.

22 In all but one of the proteins of the data base the resolution was better than 2.0 Å. Lesk. A.M. Protein Architecture, (1991) Oxford University Press, p15.

23 Momany F. A., McGuire R. F., Burgess A. W., Scheraga H. A. (1975) *J. Phys. Chem.* **79,** 2361-2381.

24 Levitt M. (1992) *J. Mol. Biol.* **226,** 507-533.

25 Weiner S. J., Kollman P. A., Case D. A., Singh U. C., Ghio C., Alagona G., Profeta S. Jr., Weiner P. (1984) *J. Am. Chem. Soc.* **106,** 765-784.

26 Still W. C., Tempczyk A., Hawley R. C., Hendrickson T. (1990) *J. Am. Chem. Soc.* **112,** 6127-6129.

27 This is done using Macromodel/BatchMin .com file by adding the command FXAT after command READ and setting zeros in the x, y, and z columns.

28 Claessens M., Cutsem E. V., Lasters I., Wodak S. (1989) *Protein Engineering* **2,** 335-345.

Chapter Four

Prediction of side-chain conformations

Summary

A model of nine proteins including side-chain atoms have been built from the known C^{α} coordinates and amino acid sequences using a Monte Carlo Protein Building Annealing (MCPBA) method. The Cartesian coordinates for the side-chain atoms were established with bond lengths and angles selected randomly from within ranges of values previously determined by analysis of fourteen protein crystal structures and with torsional angles randomly selected from -180° to 180°. A simulated annealing technique is used to generate some 300 structures with differing side-chain conformations. The atomic coordinates of the backbone atoms are fixed during the simulated annealing process. The coordinates of the side-chain atoms of the 300 low energy conformations are averaged to obtain a mean structure which is minimized with the C^{α} atoms constrained to their position in the X-ray structure using the OPLS/AMBER force field with the GB/SA water model. The r.m.s.d of the main-chain atoms (without C^{β}) compared with the corresponding crystal structures is in the range 0.20 Å to 0.64 Å with a average value of 0.45 Å. The r.m.s.d of the side-chain atoms is between 1.72 Å and 2.71 Å with a average of 2.26 Å. The r.m.s.d of all atoms is between 1.19 Å and 1.99 Å with a average of 1.61 Å. The method is insensitive to random errors in the C^{α} positions and the computational requirement is modest.

4.1. Introduction

In the past few years several groups have developed model building algorithms to construct all the atomic coordinates of a protein backbone and the side-chains from known C^{α} coordinates.[1] Two degrees of torsional freedom, ϕ and ψ, largely determine the folding patterns of the main-chain. Side-chain torsional angles similarly define side-chain conformations. These main and side-chain variables are coupled due to the importance of side-chain packing in folding stability. Early studies suggested that the side-chains adopt low energy conformations[2] often close to conformations found in free amino acids. Recent studies of a large sample of well-refined structures confirm that commonly occurring conformations are close to standard low energy rotamers of the individual amino acids.[3,4]

As with the problem of protein folding, the principal difficulty in making predictions of side-chain conformations is the enormous number of permutations possible. Three different strategies have been used to minimise this obstacle.

(i) The conformations of each residue can be reduced to a limited number of allowed conformations. This method has been useful for defining conformations that can be packed into a given spatial region.[5,6]

(ii) Side-chain conformations can be built by either molecular dynamics or simulated annealing techniques. The structures are subjected to energy minimisation. These techniques are computer intensive. For example molecular dynamics has been used[7] to generate conformers from model structures built using Scheraga's algorithm (Appendix 1) from C^{α} coordinates for α-lytic protease, troponin and flavodoxin. For these three proteins the main-chain r.m.s.d range from 0.19 - 0.49 Å, and the all atom (backbone and side-chain) r.m.s.d from 1.24 - 1.68 Å. Specifically the r.m.s.d, for 3FXN is 0.49 Å for main-chain atoms and 1.64 Å for all atoms.

Lee and Subbiah[8] generate side-chain conformers by varying side-chain torsion angles in increments of 10° during simulated annealing. The van der Waals interactions between side-chain atoms and main-chain atoms were evaluated for each rotamer and for nine proteins. The average side-chain r.m.s.d was 1.77 Å. (1Å = 0.1 nm).[9]

Recently a modelling algorithm (PROGEN) utilising an optimal geometry parameter database for the positioning of atoms of each amino acid from C^{α} coordinates has been described.[10] Subsequent molecular dynamics with the backbone atoms fixed is used to generate other conformations. The method first generates the backbone and C^{β} atoms. The side-chains are rotated about the various torsional angles and molecular dynamic simulations and energy minimisation used to place atoms in positions of minimum contacts. For sixty structures the r.m.s.d for the main-chain atoms ranged from 0.29 - 0.76 Å with an average value of 0.53 Å. For all non-hydrogen backbone and side-chain atoms the range in r.m.s.d was between 1.44 - 1.93 Å.

(iii) The main-chain and side-chain conformations can be generated by segment to segment matching with known proteins in a data base. Energy minimisation techniques have been used to refine the side-chain conformations. The side-chain conformations[11] of flavodoxin (3FXN), a commonly employed standard for such studies, have been predicted from the C^{α} coordinates with an overall side-chain r.m.s.d of 2.4 Å. Structures generated by this method[12] have been used to develop electron density maps[13] as an aid to X-ray structure analysis.

In another segment match modelling approach (SMM)[14] the target structure is broken into short segments defined only by C^{α} coordinates and the sequence. The segments were matched in the protein data base and the segment coordinates fitted into the growing target structure. The procedure can be repeated to give a number of independently built models. The structures were averaged to give a mean structure, often with unrealistic bond lengths and angles which was subjected to energy minimisation to bring appropriate corrections to bond lengths and angles. The main strength of this strategy, is its ability to circumvent the number of permutations of side-chain conformations. However the limitation of the method is the extensive and inclusive protein data base required.

We have previously developed a Monte Carlo Protein Building method (MCPB) to generate the main-chain atomic and C^{β} atoms coordinates from C^{α} coordinates.[15] We have extended this method, modified to include simulated annealing[16] and referred to as the Monte Carlo Protein Building Annealing method (MCPBA) to generate a battery of

conformers to the generation of side-chain atoms. The performance of the method for the generation of side-chain conformations is assessed.

Our attention is focused toward modelling the coiled-coil protein structures found in wool protein. Few X-ray structures of proteins that contain coiled-coiled segments exist in the PDB file, and therefore segment matching techniques are not yet appropriate. The MCPB-anneal (MCPBA) technique requires less computing time than for a systematic search of side-chain conformations[5] or molecular dynamics strategies[7] for defining side-chain conformations. These latter methods would be prohibitive for proteins of the size in wool. We have applied the MCPBA technique to the side-chain packing problem for nine proteins to assess its performance in the prediction of the side-chain atomic coordinates.

The method faces two distinct challenges. Firstly, it is unclear the extent to which packing energy is a predictor of protein structure since even relatively sophisticated potential functions have been found to be poor predictors of protein structure.[17] For this reason, and due to the complexity of the problem we have in the first instance limited the computation of side-chain conformational energies to evaluation of van der Waal interactions. Secondly, any method must overcome the combinatorial complexity to find well-packed conformations.

4.2. Computational methods

The atomic coordinates of the backbone atoms (including C^{β}) are defined by the Monte Carlo Protein Building method (see chapter three) where dihedral angles are selected from a data base arrived at from a statistical analysis of the relationship between dihedral angles of the backbone and C^{α} coordinates for a protein data base. The averaged coordinates from ten backbone models of a protein were used to define a mean structure that was refined by energy minimisation (EM) generally using the Amber/OPLS force field with GB/SA water model.

4.2.1. Threshold energy of the side-chain

The coordinates of the C^{γ} atom of the first side-chain from the N-termini are generated from the C^{β}, C^{α} and N coordinates of the first amino acid residue of the

backbone using Scheraga's algorithm.[18] The bond length and bond angles associated with C^{γ} are taken by random selection from literature values.[19] Statistical analysis of side-chain conformations from X-ray structures have shown[4] the side-chain torsional angles are general χ = 60 (g-), 180 (t) or 300 (g+).[3,20,21] but criterion for predicting the value are not known. For these reasons, except for the proline residue, we have, unlike other methods[23] not limited the selection of torsional angles from preferred ranges but made random choices from -180° to 180°. The subsequent atomic positions of the side-chain are defined in a similar way from the internal coordinates of atoms C^{α}, C^{β}, C^{γ} and so on. The proline residue is special because little variation in the side-chain torsional angles are observed.[22] In the crystal structures there are three populated side-chain conformations for proline, defined as up, down and planar.[23] The ranges of values of the torsion angles in these three conformations are normally within the ranges χ_1 = 10° - 40° and χ_2 = 10° - 15° and values of the torsional angle for proline are therefore randomly selected from within these ranges.

In order to obtain the energy associated with a side-chain conformation the van der Waals 6-12 potential energy[24] is determined using the non-hydrogen energy parameters of the force fields ECEPP/2, AMBER and OPLS developed by Scheraga,[23,25] Kollman[26] and Jorgensen[27] respectively. For the models of the proteins 1CRN, 1ROP, 3FXN and 2ZTA the OPLS van der Waals force field gave the lowest r.m.s.d of all non-hydrogen atoms after energy minimisation. We have therefore used this force field in most subsequent studies.[28] The van der Waals interaction energy is successively calculated for each side-chain atom of a residue in associated with that residue backbone atoms (N, C^{α}, C', O) and all atoms of the previous four amino acid residues. (For the first few amino acids it is obviously not possible to consider four amino acids but only those that are present). This gives a van der Waals interaction energy for each atom of the side-chain. For any conformation of the side-chain the energies of each of the side-chain atoms obtained in this way are added together and this energy is stored as the energy for the side-chain, for that conformation and the environment of the specific conformations of the preceding four amino acids. The next side-chain is then constructed and energy of the side-chain assessed as above. The process is continued to the last residue side-chain. The

process is repeated to generate three hundred conformers of the protein have been generated. The average value of the conformational van der Waals energy for each residue of the side-chain for the 300 structures is determined and is used as a threshold value for this particular residue side-chain in subsequent simulated annealing.

4.2.2 The simulated annealing protocol.

With the backbone atoms fixed, structures with varying side-chain conformations are generated as above. These can be considered as structures **301**, **301**, **302,** etc. The first structure (i.e. **301**) is used as a mother conformer. A second structure **302** is built with each side-chain being added sequentially in the following way. The van der Waals energy of each side-chain with respect to the amino acid that it is attached to and to the preceding four amino acids is assessed as above. If the conformational energy of the side-chain is lower than or equal to the threshold energy of that side-chain in the average of the 300 structures above then the atomic coordinates of this side-chain in the new conformer are retained. The next side-chain is built and the van der Waals energy assessed. If the van der Waals energy of this side-chain with the amino acid to which it is attached along with the proceeding four amino acids and side-chains is above the threshold energy then the atomic coordinates of this side-chain are discarded and the atomic coordinates of the side-chain in the mother conformer are copied into the side-chain position of the new conformer. This process is repeated for the next side-chain and so on until all the side-chains have been assigned a conformation and the structure of **302** is then complete. The total van der Waals energy interaction of this conformer is calculated.

A simulated annealing criterion is used to determine if this new structure **302** is accepted or discarded. If the total van der Waals energy of **302** is less than a "predetermined arbitrary large value", it is accepted and the atomic coordinates kept. The energy of this structure now replaces the energy of the "predetermined large value energy". The next structure **303** is generated in the same way as **302** and if the conformational energy of **303** is not less than that of the previous structure **302**, an acceptance criterion is invoked. If the value of $\exp(-\Delta E/T)$ ($\Delta E = E_{new}(\mathbf{303}) - E_{old}(\mathbf{302})$ where T is the annealing temperature) is greater than or equal to r i.e. $r \leq$

exp(-ΔE/T), where r is an uniform random number (0 < r < 1), the structure is accepted and the energy E_{old}(**302**) of the previous structure is replaced by the energy E_{new}(**303**) of the new structure. If r > exp(-ΔE/T), then the new structure is rejected. This procedure prevents the conformational search being locked into a local minimum. The annealing temperature is gradually lowered from 1500°C to ca 5°- 50°C. A series of conformers of the modelled protein are generated in this way.

As each structure is accepted on energy criteria the r.m.s.d of the main-chain, the side-chain and all atoms are compared with the corresponding X-ray structure. The structure number, energy and r.m.s.d data is recorded in an .acp file though at this stage only the coordinates of the lowest energy structure are kept. In the simulated annealing procedure, the ratio of the number of accepted conformational energies to the total number of conformations generated is represented by P which has a value within the range 0 and 1. At high temperatures, almost all the energies of the conformations generated are accepted since T is greater than ΔE, ((ΔE/T) ≅ 0) and exp(-ΔE/T) is close to 1. As the temperature is lowered the proportion of the conformers accepted reduces (P gradually lowers). If the ratio P falls too quickly, the simulated annealing process can be trapped in a local energy minimum. The annealing temperature T governs the rate of change of P. The compromise is that the more slowly P falls the more computer time is required. In practice, T is lowered in discrete steps. The starting temperature for annealing is given as T_S [8] and a temperature factor, g, where g is between 1 and 0.1, is included in the calculation to control the rate of temperature fall. If the energies of two consecutively generated conformers are accepted the temperature is reduced by multiplying the temperature by the value g. For this work a value of 0.98 was found appropriate.[8] As the simulated annealing proceeds, the frequency of conformer rejection increases and the ratio P gradually decreases. When P = 0.5 half of the generated conformations have had energies that have been recorded in the .acp file though the coordinates at this stage are not collected. For values of P lower than 0.5 the coordinates of each structure that is accepted on energy criteria are retained in the .scg file. For practical reasons of computer storage space, we have chosen to collect the coordinates of a maximum of 300

conformers at any time. As each new structure over 300 is kept then the first of the three hundred structures are discarded and so on.

The termination of the annealing procedure is set by one of three conditions: (i) a defined number of accepted conformations is met. For our studies the number was set at 20,000, (ii) a predetermined number of consecutively rejected conformers is met. For this work the value was set at 1,000, (iii) or when P = 0.1. These criteria, the force field used and the starting temperature T_S can be defined in the *crystal.com* file. The temperature at which the simulated annealing stops is T-finish (T_f).

4.2.3. The mean structure of the full atom models

The averaging of atomic coordinates of several modelled structures of a protein backbone has proved useful in generating best structure coordinates.[14] The average of ten backbone and C^β models generated by the MCPB method for each of twenty four proteins gave an r.m.s.d of 0.43 Å.[29] For the ensemble of 300 structures collected by simulated annealing we have examined the effect of averaging the main-chain and side-chain coordinates after the simulated annealing process is terminated. The r.m.s.d of the mean structure is significantly smaller than that for any individual conformer, however the bond lengths and angles of the mean structure are somewhat unrealistic. Bond lengths of the side-chain are shorter than expected and aromatic rings distort.[30] The problem arises because of the large differences in side-chain conformations and is overcome by energy minimisation when the angles and lengths are brought to realistic values.[14] These values are corrected by energy minimisation of the mean structure carried using the *Macromodel/Batchmin* package.[31]

4.2.4. Energy minimisation of the mean structure

An average structure obtained after simulated annealing procedure without hydrogen atoms for each of the twenty-four proteins was minimised with *Macromodel/Batchmin* using the AMBER/OPLS (non-hydrogen) force field[26] and GB/SA water model.[32] The C^α atoms of the model structure of the proteins are restrained

to their initial positions as defined by the X-ray coordinates.[33] Energy minimisation is limited to 300 steps.

4.2.5. Computer program

The program *crystal,* written in FORTRAN 77 to run on the IBM RS/6000 generates the backbone structure using the MCPB method and builds the side-chain conformations by use of the Scheraga algorithm and encapsulated the simulated annealing protocol and screens the structures as appropriate. In the program, two steps are involved. First ten independent models of the backbone and C^{β} atoms are generated with different random seed numbers by the MCPB method. The atomic coordinates of the ten structures are averaged to obtain a mean structure. Secondly the Scheraga algorithm is used to generate a conformation of each side-chain and this is followed by the simulated annealing procedure to generate the side-chain conformations about the mean backbone structure using the MCPBA procedure. Generally some 5,000 conformers of the protein are generated with the atomic coordinates of 300 conformations collected at any time during the annealing procedure. The program *aver*, also written in FORTRAN 77 gives the mean structure by calculating the average atomic coordinates of the collected conformations. Energy refinement of the mean structure is carried out using Macromodel/Batchmin. The computer requirements of these procedures are modest. For proteins with ca 200 residues, it takes ca. 2 hours for simulated annealing processing. 3FXN with 138 residues takes 92 minutes. Energy minimisation of the mean structure (300 iterations) takes ca. 2 hours for a protein with 200 residues. Thus, the total computing time required to model such a protein is about 4 hours.

4.3. Results and discussion

The MCPBA method has been tested on nine proteins (see Table 4.1) selected from the Brookhaven Protein Data Bank[34], which have a resolution of 2.5 Å or better.[35] The proteins contain between 46 and 494 amino acid residues and the total number of the residues in the nine proteins is 1338. Three of the proteins 1TIM, 2CCY and 2ZTA, have

two chains or sub-units. 2ZTA is a coiled-coil protein with two parallel chains[36] and 1ROP is a coiled-coil where one chain folds into an anti-parallel coiled-coil structure.[37]

Table 4.1. The proteins used in this work

Protein Name	Code	Number of Residues
Crambin	1CRN	46
COL*E Rop	1ROP	56
Isomerase	1TIM	494
Steroid Binding	1UTG	70
Cytochrome C	2CCY	254
Leucine Zipper	2ZTA	62
Flavodoxin	3FXN	138
Trypsin Inhibitor	4PTI	58
Troponin C	4TNC	160

4.3.1. The r.m.s.d. of full atom model

The atomic coordinates of the side-chains of the nine proteins have been generated using the MCPBA method. The r.m.s.d of side-chain, the backbone atoms and all non hydrogen atoms of the mean structure after energy minimisation are compared with the X-ray structures in Table 3.2 of chapter 3. The models constructed from the known C^{α} coordinates by the MCPBA method are remarkably accurate. After energy minimisation, the overall r.m.s.d of the protein models is 0.45 Å for main-chain atoms (excluding C^{β}), 2.25 Å for side-chain atoms and 1.61 Å for all non-hydrogen atoms. The accuracy varies with the protein. The r.m.s.d of all atoms is lowest for 1CRN (1.19 Å) and greatest for 4PTI (1.99 Å). The r.m.s.d of the mean structure including main-chain and side-chain atoms is better than the average r.m.s.d of each of the collected conformations.

A maximum difference of ca. 1Å between the average r.m.s.d of the collected 300 structures and the r.m.s.d of the mean structure, obtained from the 300 structures and before energy minimisation occurs for the side-chain of protein 4PTI. Energy minimisation of a mean structure can slightly increase the r.m.s.d of the side-chain and all atoms, however distortions on bond lengths and angles of a mean structure are corrected.

The r.m.s.d of the backbone atoms show less variation between the different protein structures than do the side-chain or all atom r.m.s.d's after energy minimisation. The average value of the r.m.s.d of the main-chain is 0.45 Å, which is comparable to that previously reported (0.43 Å) in chapter three. The protein 3FXN after energy minimisation on the mean structure, with the side-chains attached and C^{α} carbons fixed in the position defined by the X-ray structure, has a r.m.s.d for the main-chain atoms of 0.57 Å (Table 4.2).[38]

Table 4.2. The r.m.s.d of protein model compared with the crystal structure.

		R.m.s.d. (Å)				
Protein		Simulated annealing			Mean[e]	EM[f]
	Nres[a]	Worst[b]	Best[c]	Ave[d]		
Main-chain						
1CRN	46	0.96	0.72	0.84	0.64	0.52
1ROP	56	0.52	0.43	0.46	0.36	0.33
1TIM	494	0.86	0.76	0.79	0.58	0.62
1UTG	70	0.66	0.61	0.53	0.38	0.30
2CCY	254	0.69	0.63	0.66	0.48	0.48
2ZTA	62	0.48	0.37	0.42	0.23	0.21
3FXN	138	0.81	0.71	0.77	0.53	0.57
4PTI	58	1.03	0.72	0.89	0.61	0.64
4TNC	160	0.66	0.56	0.61	0.39	0.40
Overall[g]						0.45
Side-chain						
1CRN	46	3.25	1.81	2.51	1.79	1.72
1ROP	56	3.19	2.44	2.82	1.93	2.04
1TIM	494	3.24	2.79	3.05	2.21	2.36
1UTG	70	3.12	2.34	2.76	1.96	2.04
2CCY	254	3.19	2.63	2.96	2.13	2.40
2ZTA	62	3.62	2.56	3.15	2.21	2.57
3FXN	138	3.27	2.68	3.01	2.20	2.15
4PTI	58	4.18	2.98	3.62	2.68	2.71
4TNC	160	3.19	2.54	2.90	2.07	2.37
Overall						2.25

All atoms						
1CRN	46	2.16	1.27	1.70	1.07	1.19
1ROP	56	2.26	1.73	1.99	1.38	1.45
1TIM	494	2.25	1.95	2.13	1.57	1.68
1UTG	70	2.19	1.66	1.95	1.34	1.47
2CCY	254	2.16	1.79	2.01	1.47	1.64
2ZTA	62	2.62	1.85	2.28	1.62	1.89
3FXN	138	2.30	1.90	2.12	1.58	1.55
4PTI	58	2.96	2.13	2.57	1.94	1.99
4TNC	160	2.24	1.80	2.04	1.47	1.67
Overall						1.61

a Nres is the number of residues in the protein.

b Worst: the highest r.m.s.d. in the collected structures compared with the X-ray structure.

c Best: the lowest r.m.s.d. in the collected structures compared with the X-ray structure.

d Ave: the average value of r.m.s.d for all collected structures in the simulated annealing.

e Mean: the r.m.s.d. of the mean structure before energy minimisation.

f EM: the r.m.s.d. of the mean structure after energy minimisation.

g The overall r.m.s.d. corresponding to values for the nine proteins.

4.3.2 The r.m.s.d. of amino acid residues

An examination of the r.m.s.d for individual amino acids in the nine energy minimised modelled structures are shown in Table 4.3. Bond lengths and angles are close to standard values.[39]

Table 4.3. R.m.s.d of the amino acids in the nine proteins.

		r.m.s.d. (Å)		
Amino acid	number	backbone[a]	side-chain	total[d]
Ala	151	0.38	0.35	0.38
Arg	46	0.44	3.08	2.47
Asn	41	0.51	2.08	1.53
Asp	78	0.49	2.26	1.64

Gln	46	0.32	2.59	1.03
Cys	32	0.41	1.53	0.96
Glu	123	0.39	2.37	1.79
Gly	114	0.72	--b	--b
Ile	76	0.39	1.83	1.33
His	22	0.33	2.22	1.76
Leu	108	0.38	1.77	1.29
Met	35	0.51	2.19	1.62
Lys	118	0.41	2.59	1.92
Phe	48	0.45	2.86	2.30
Pro	46	0.50	0.56	0.55
Ser	60	0.41	1.35	0.88
Thr	62	0.42	1.57	1.12
Trp	19	0.43	3.32	2.83
Tyr	21	0.40	2.87	2.37
Val	80	0.30	1.26	0.77
NAM[c]	12	0.42	--b	--b
Overall	1338	0.43	2.04	1.41

[a] Backbone atoms without C^{β}, [b] No side-chain for this amino acid, [c] NAM represents the capped group for the last residue of the protein. There are twelve chains or sub-units in the nine proteins. 2ZTA, 2CCY and 1TIM have two chains or sub-units. There are a total of twelve capped groups. d The r.m.s.d. of the structure including all backbone and side-chain non hydrogen atoms.

For most amino acid main-chain atoms the r.m.s.d is ca. 0.4 Å. The r.m.s.d is poorest for glycine (Gly) (0.71 Å) which is not constrained by a side-chain. An analysis of r.m.s.d of main-chain and C^{β} atoms has been reported previously (see chapter three). The r.m.s.d's of the side-chain atoms are below 1 Å for Proline (Pro) and Alanine (Ala), and below 2 Å for Cysteine (Cys), Valine (Val), Leucine (Leu), Isoleucine (Ile), Serine (Ser) and Therine (Thr). The polar and charged amino acid residues have larger r.m.s.d's than for non-polar residues. The polar amino acid residues and those containing aromatic groups have r.m.s.d's of the side-chain atoms greater than 2.0 Å. The largest r.m.s.d is for the polar amino acids Tryptophan (Trp), followed by Arginine (Arg) which are more

difficult to model because their conformations are most affected by solvent and crystal contacts.[14,10]

Structural diagrams of the amino acid residues with aromatic side-chains overlaid with the corresponding atoms from the X-ray structures for 3FXN and 2CCY are shown in Figure 4.1 and Figure 4.2.

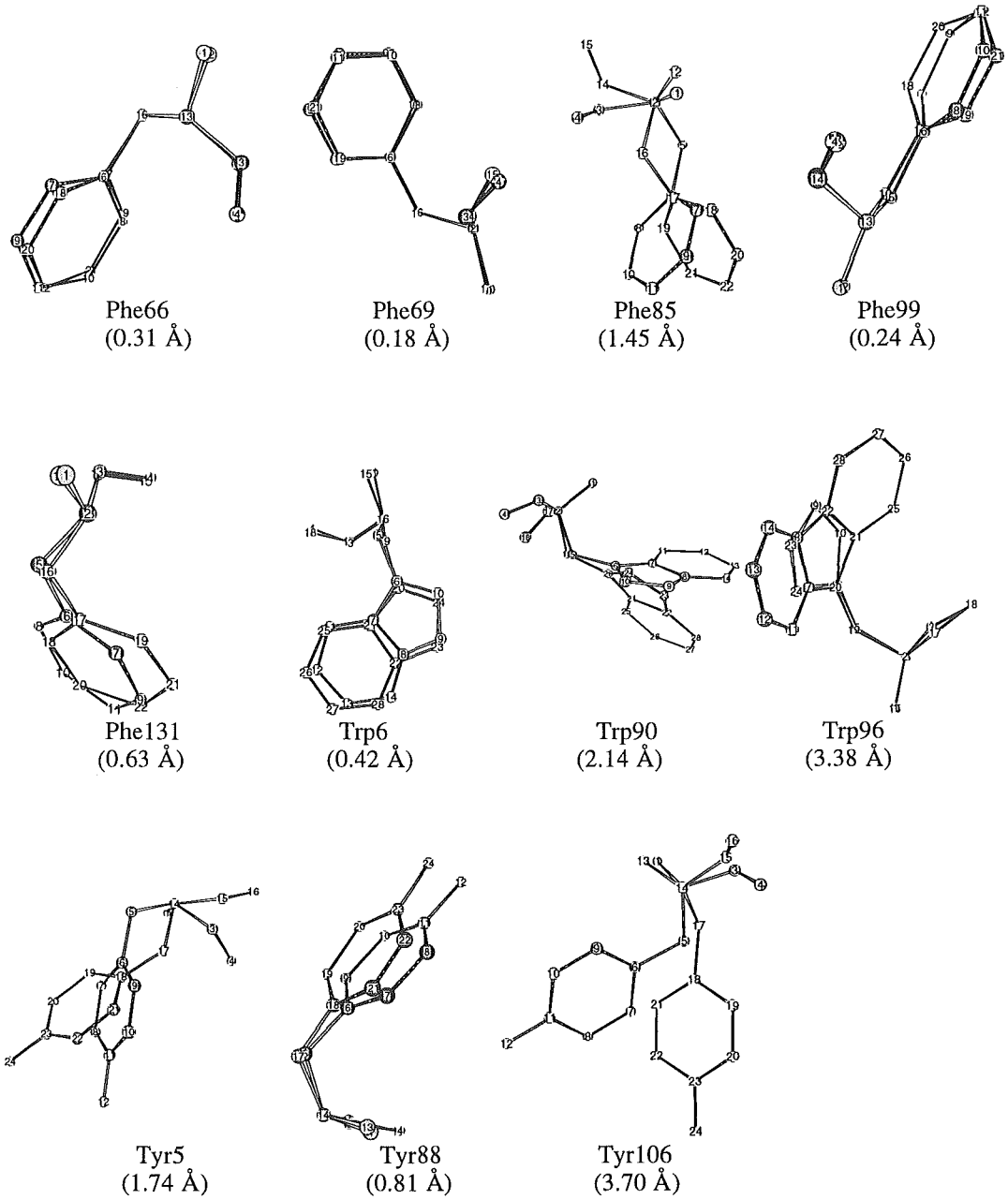
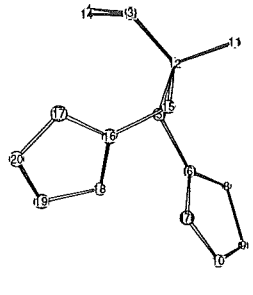
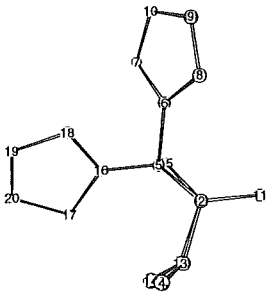


Figure 4.1. Structural diagrams of residues with aromatic side-chain: X-ray structures and the corresponding modelled structure built by the MCPB and MCPBA methods for the C^{α} coordinates of 3FXN. The backbone atoms (N, C^{α}, C', O) and the side-chain atoms are drawn for each residue. The numbering of atoms in each residue

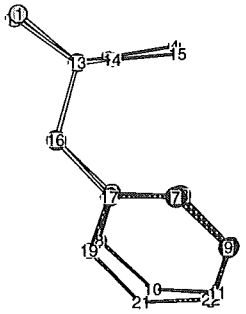
begins at the nitrogen of the backbone. The X-ray structure have atom numbers is less than 14 for Trp, 12 for Tyr, and 11 for Phe, and the modelled structure atoms numbers larger than that values respectively. The r.m.s.d. of the residue side-chain (including C^{β}) compared to the corresponding X-ray structures are given in parentheses. The Structural diagrams were made on an Apple Macintosh by using Chem3D Plus software.



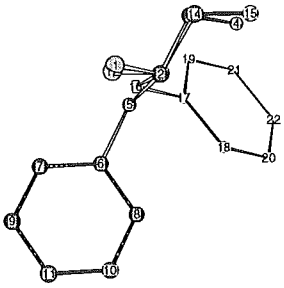
His121
(3.77 Å)



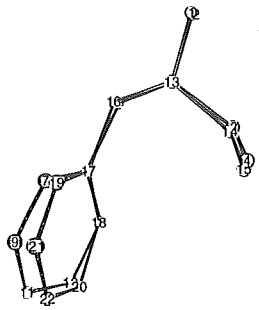
His248
(3.91 Å)



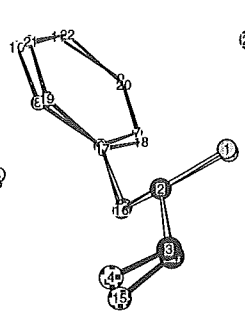
Phe28
(0.25 Å)



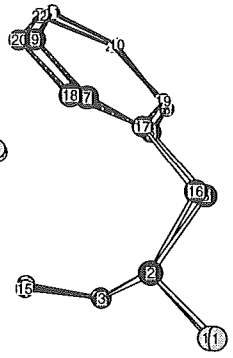
Phe74
(5.25 Å)



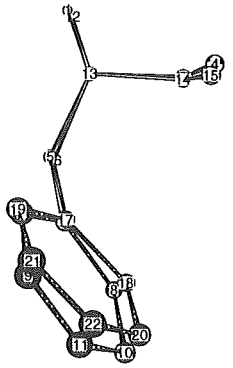
Phe81
(0.32 Å)



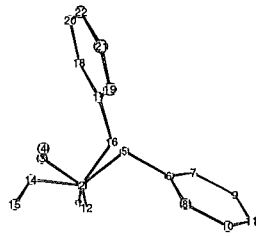
Phe124
(0.17 Å)



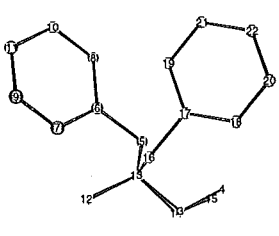
Phe155
(0.22 Å)



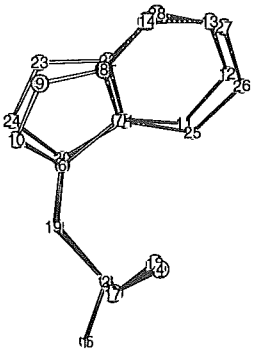
Phe208
(0.55 Å)



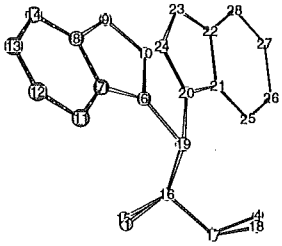
Phe251
(5.54 Å)



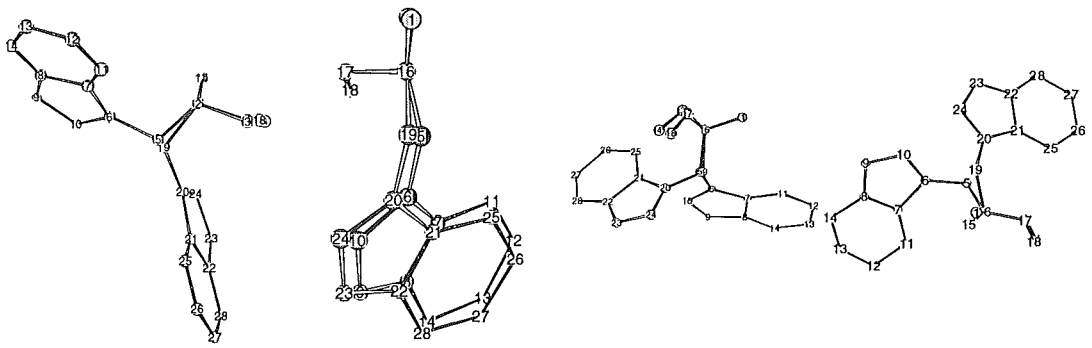
Phe201
(4.39 Å)



Trp22
(0.44 Å)



Trp67
(5.39 Å)



Trp85 (6.79 Å) Trp149 (0.51 Å) Trp184 (5.59 Å) Trp212 (6.38 Å)

Figure 4.2. Structural diagrams of the aromatic residues; X-ray structures and the corresponding modelled structure built by MCPB and MCPBA methods using the C^{α} coordinates of 2CCY. The backbone atoms (N, C^{α}, C', O) and the side-chain atoms are drawn for each residue. The numbering of atoms in each residue begins at the nitrogen of the backbone. The X-ray structures have atom numbers less than 14 for Trp, 11 for Phe and 10 for His, and the modelled structure have different numbers larger than these values respectively. The r.m.s.d. of the residue side-chain (including C^{β}) compared to the corresponding X-ray structures are given in parentheses. The molecular stereo views were made on an Apple Macintosh by using Chem3D Plus software.

For 3FXN, the residues Trp6, Phe69, Phe66, Tyr88, Phe99, Phe85 and Phe131 are close to conformation in crystal structure. Residues Tyr5, Tyr106, Trp90 and Trp95 are packed differently in the modelled to the crystal structures. The r.m.s.d. of all side-chain atoms is 2.1 Å.

For 2CCY, the side-chains Phe28, Phe81, Phe124, Phe155, Phe208, Trp22, and Trp149 are close to conformation in the crystal. The side-chains of His121, His248, Phe74, Trp184, Trp57, Trp85, Phe251, Trp212 and Phe201 are packed differently in the model compared with the crystal structure. The largest r.m.s.d. is for the side-chain Trp85 with the value of 6.79 Å. The r.m.s.d. of all side-chain atoms is 2.4 Å.

If the average r.m.s.d of the side-chain atoms compared with the X-ray structure exceeds 2.0 Å, the side-chain conformation is considered to be incorrect. This criterion has been introduced by Levitt[14] in his segment matching method. For the nine proteins in this study the number of side-chains outside this criteria for each amino acids is given in Table 4.4. Polar side-chains have an error rate of 31% which is lower than for non-polar

side-chains. The error rate for charged side-chains is high and greatest for Arg, consistent with interaction between polar side-chain and solvent having an important influence on conformation. Non-polar side-chains have an error rate of 18% with the exception of methionine (Met) with a high error rate of 48%.

Table 4.4. Ratio of the number of side-chains for the different amino acids which are outside the 2 Å criteria

Amino acid	1CRN	1ROP	1TIM	1UTG	2CCY	2ZTA	3FXN	4PTI	4TNC	Overall
Special										9/204
NAM[a]	0/1	0/1	0/2	0/1	0/2	0/2	0/1	0/1	0/1	0/12
Gly	0/4	0/0	0/54	0/3	0/20	0/0	0/14	0/6	0/13	0/114
Pro	0/5	0/0	0/14	0/5	0/14	0/0	0/3	0/4	0/1	0/46
Cys	1/6	0/2	2/8	2/2	1/4	0/0	0/3	3/6	0/1	9/32
Non-polar										84/450
Ala	0/5	0/6	0/56	0/2	0/56	0/2	0/6	0/5	0/13	0/151
Val	0/2	0/0	11/46	1/3	1/6	1/6	2/10	0/1	1/6	17/80
Ile	1/5	0/2	11/34	1/4	0/4	0/0	8/14	0/2	1/11	22/76
Leu	0/1	3/9	14/34	3/8	3/24	0/12	3/8	0/2	2/10	28/108
Met	0/0	1/2	3/4	2/4	2/6	1/2	3/5	0/1	5/11	17/35
Aromatic										57/110
His	0/0	0/2	7/14	1/1	2/2	0/2	0/0	0/0	1/1	11/22
Phe	0/1	0/1	12/16	1/2	3/8	0/0	0/5	1/4	7/11	24/48
Tyr	0/2	0/1	5/8	0/1	0/0	2/2	1/3	1/4	0/0	9/21
Trp	0/0	0/0	7/10	0/0	4/6	0/0	2/3	0/0	0/0	13/19
Polar										66/210
Ser	0/2	0/3	3/24	0/5	0/10	0/2	0/8	0/1	0/5	3/60
Thr	2/6	1/4	5/20	0/6	1/12	0/0	0/5	0/3	3/6	12/62
Asn	1/3	0/2	10/12	0/2	1/4	1/4	6/8	1/3	4/4	24/42
Gln	0/0	3/3	10/18	2/2	7/14	2/2	1/2	1/1	3/4	27/46
Charged										237/365
Asp	1/1	2/5	11/26	2/4	3/8	2/2	7/9	2/2	15/21	45/78
Glu	0/1	4/6	19/34	5/6	12/20	6/10	11/19	2/2	18/25	77/123
Lys	0/0	1/3	30/44	6/7	20/30	6/10	7/10	3/4	6/10	79/118
Arg	2/2	1/4	14/16	1/2	3/4	4/4	2/2	5/6	6/6	36/46
All	8/46	17/56	176/494	28/70	66/254	25/62	54/138	19/58	73/160	453/1338

%	17.3	30.4	35.6	40.0	26.0	40.1	39.1	32.7	45.6	33.5

[a] NAM is the capped group of an end-chain.

Overall 33% of the side-chains fall outside the 2 Å criteria. When charged side-chains are excluded the error rate falls to 22%. For 1CRN the overall error rate is lowest (17.3%) and is highest for 4TNC (45.6%). The "special" group of the amino acids which includes capped amino acids, glycine, proline and cysteine have an error rate of 4%.

The r.m.s.d for the main-chain, side-chain and all atoms for the different motif regions of secondary structure in each protein are listed in Table 4.5.

Table 4.5. R.m.s.d in the secondary structure region of the model in the nine proteins

	Motif											
Protein	α -Helix			β -Sheet			β - Turn			Random coil		
	B[a]	S[c]	T[d]	B	S	T	B	S	T	B	S	T
1CRN	0.30	1.94	1.32	0.70	1.75	1.23	0.33	1.87	1.25	0.69	0.87	0.75
1ROP	0.20	2.04	1.44	--[e]	–	–	–	–	–	1.99	3.79	3.03
1TIM	0.59	2.38	1.71	0.56	2.03	1.44	--[e]	–	–	0.64	2.40	1.70
1UTG	0.18	2.27	1.61	--[e]	–	–	0.39	3.42	2.61	0.54	1.67	1.13
2CCY	0.35	2.44	1.67	--[e]	–	–	0.69	2.61	1.73	0.54	2.39	1.64
2ZTA	0.23	2.60	1.89	--[e]	–	–	–	–	–	–	–	–
3FXN	0.23	2.33	1.65	0.45	1.94	1.35	0.91	2.13	1.62	0.71	2.21	1.56
4PTI	0.20	2.60	1.81	0.40	2.28	1.73	--[e]	–	–	0.78	3.11	2.16
4TNC	0.27	2.38	1.71	--[e]	–	–	–	–	–	0.55	2.60	1.80
Overall	0.28	2.33	1.64	0.53	2.00	1.44	0.58	2.51	1.80	0.81	2.38	1.72

[a] B; the r.m.s.d. of the backbone atoms, [c]S; the r.m.s.d. of the side-chain atoms, [d]T; the r.m.s.d. of all non-hydrogen atoms. [e] no motif is assigned in the PDB. The motifs, α-helix, β-sheet and β-turn, are as defined in the PDB. Residues which have no definition in the PDB are taken as a random coil motif.

For the backbone atoms of α-helix motif the r.m.s.d (0.28 Å) is smallest. For β-sheet 0.53 Å, β-turn (0.58 Å) and random coil motifs (0.81Å). For side-chains of α-helical motif the r.m.s.d (2.33 Å) is greater than for β-sheet (2.00 Å). The r.m.s.d of the side-chain of CRN is 1.94 Å in the α-helical region and 1.75 Å in the β-sheet region. For FXN the r.m.s.d of the side-chain is 2.33 Å in the α-helical region and 1.94 Å in the β-sheet region. In PTI, 2.60 Å in the α-helical region and 2.28 Å in the β-sheet region. For TIM, the r.m.s.d is 2.38 Å in the α-helical region and 2.03 Å in the β-sheet region. The r.m.s.d of the side-chain conformation in the β-sheet region is significantly better than that in the α-helical region. The side-chain conformation in β-sheet regions is modelled better than in α-helical regions by the MCPBA method. In the β-turn region the average r.m.s.d (2.51 Å) of the side-chain conformations is greatest. The random coil region the r.m.s.d of the side-chain is 2.38 Å close to that in α-helix motif (2.33 Å).

4.3.3. Molecular surface areas and volumes

The solvent-accessible surface (SAS) area and volumes and van der Waals (VDW) surface areas and volumes of both the modelled (MCPB/MCPBA) and X-ray structures were calculated with the *Gepol* program using the method of Pascual-Ahuir et al.[40] The SAS areas and volumes of the proteins are measured using a water molecule modelled as a spherical probe with a radius of 1.4 Å (Table 4.6).

Table 4.6. The SAS and VDW surface areas and volumes of the model and the X-ray structure.

	Surface area (Å^2)				Volume (Å^3)			
	SAS		VDW		SAS		VDW	
Protein	Model	X-ray	Model	X-ray	Model	X-ray	Model	X-ray
1CRN	3254	3040	4304	4241	9178	8912	4180	4166
1ROP	4682	4441	6015	5905	12816	12537	5671	5644
1TIM	20675	19703	46249	45629	93624	90592	35352	47688
1UTG	5558	5127	7448	7260	16026	15305	7127	7135
2CCY	13580	13172	24989	24703	49451	48226	24042	24002
2ZTA	5518	4831	7215	7023	15451	14515	6880	6843

3FXN	7508	6974	14163	13908	27738	26847	13725	13684
4PTI	4246	3953	6061	5899	12683	12150	5773	5740
4TNC	11351	9896	34836	32903	16805	16486	16089	16075

The SAS and VDW areas of the model structures are generally larger by between 200 to 500 $Å^2$ (3-15%) and 60 to 200 $Å^2$ (1-6%) respectively than for the corresponding X-ray structure. The SAS and VDW volumes of the model structures are similarly larger than the values of the corresponding X-ray structures by 260 to 1000 $Å^3$ (4-10%). The VDW volumes of the model structures are closer to the values of the corresponding X-ray structure. The average VDW volumes compared with the X ray structures are 8 to 200 $Å^3$ larger. For example, for UTG, the VDW volume (7127 $Å^3$) of the model structure is nearly identical to its X-ray VDW volume (7135 $Å^3$). For 3FXN, used as an example to consider differences of SAS volumes and VDW surface areas between the model and X-ray structure, a plot of the SAS area and VDW area for individual amino acid residues with a surface probe of the size of a water molecule (r = 1.4 Å) is shown in Figures 4.3 and 4.4. The SAS areas for the residues have similar values between the model structure (dashed line) and the corresponding X-ray structure (solid line) (Figure 4.3). The VDW areas of the individual residues of the model structure (dashed line) are closer to those calculated for the X-ray structures (solid line)[41] than the SAS areas.

Figure 4.3. A plot of the SAS areas for the residues in the model and X-ray structure of 3FXN.

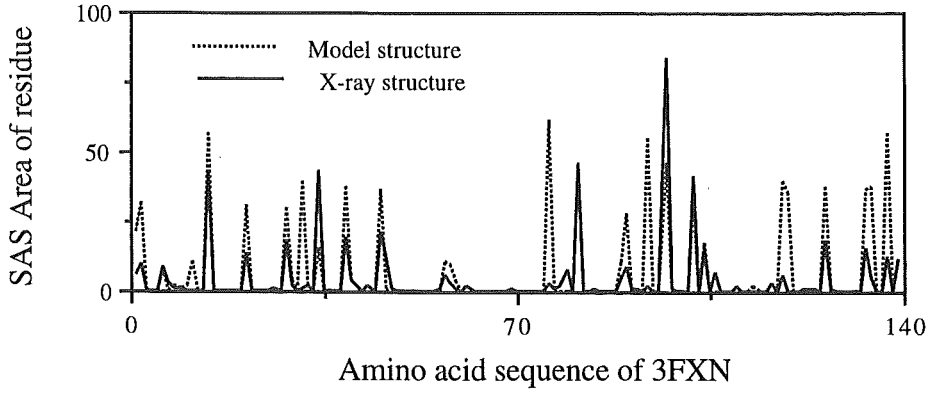
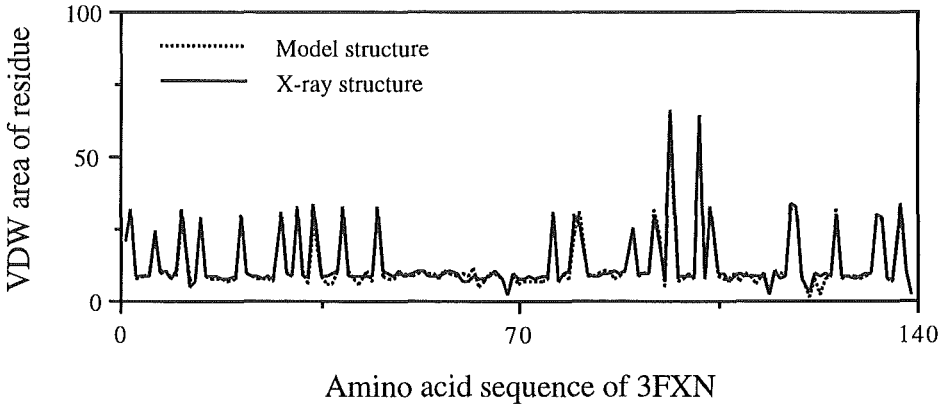


Figure 4.4. A plot of the VDW areas for the residues in the model and X-ray structures of 3FXN.



In previous studies [14,10] the r.m.s.d's of the core or buried residues have been shown to be different from that of the exposed residues. We have examined the r.m.s.d of the SAS areas of the core and of the exposed residues of the nine proteins under study. A residue is defined as a core residue if 5% or less of its potential accessible surface area is available for contact with solvent.[42] To obtain the potential solvent accessible surface area of each particular amino acid, the global minimum conformation of each of the twenty amino acids was minimised with Macromodel with the Amber/OPLS force field and GB/SA solvent model and the solvent accessible surface area calculated by using the program GEPOL. The r.m.s.d of the side-chain atoms were compared with the crystal structure the solvent accessible surface areas of the side-chain calculated. If the solvent accessible surface area of the side-chain in the model is equal to or less 5% than its potential solvent accessible area for that residue, the r.m.s.d of this side-chain residue is collected in the group of core residues. If not, the side-chain residue is considered to be exposed. Table 4.7 gives the r.m.s.d of the core and exposed residues for the nine proteins.

Table 4.7. The r.m.s.d of the core and exposed residues.

	r.m.s.d.[a]			r.m.s.d.[b]		
Protein	backbone	side-chain	all atoms	backbone	side-chain	all atoms
1CRN	0.49	1.46	1.01	0.70	2.56	1.93
1ROP	0.37	1.90	1.33	0.15	2.48	1.88
1TIM	0.62	2.26	1.59	0.56	2.83	2.16
1UTG	0.24	1.99	1.39	0.50	3.05	2.25
2CCY	0.46	2.41	1.65	1.05	2.16	1.69
2ZTA	0.14	2.33	1.66	0.34	3.40	2.56
3FXN	0.56	2.01	1.44	0.67	2.71	2.06
4PTI	0.65	2.15	1.53	0.64	3.93	3.06
4TNC	0.37	2.33	1.63	0.58	2.99	2.28
Overall	0.43	2.09	1.42	0.57	2.90	2.21

a r.m.s.d. of the core residues, *b* r.m.s.d. of the exposed residues.

The average r.m.s.d of the main-chain atoms of the core residues is 0.43 Å which is less than for the exposed residues (0.57 Å) (Table 7). Similarly the r.m.s.d of the side-chain atoms of the core residues is 2.09 Å and is less than for exposed residues (2.90 Å). The average r.m.s.d. of all atoms in the core residues is 1.42 Å which is less than for the exposed residues (2.21 Å). The packing of the core residue side-chains is more efficient for the exposed residues.

4.3.4. Effect on the random statistical errors in the C^{α} coordinates

To examine the validity of using the MCPB/MCPBA method to build accurate protein models from unrefined C^{α} co-ordinates, which are in themselves imprecise, we introduced random statistical errors to the C^{α} coordinates of the nine proteins and measured the resulting r.m.s.d's of the resulting models with respect to the X-ray structures. The random error was introduced as follows: For each C^{α} atom in the data set of the protein, the product of three random numbers between -1.0 to 1.0 with a pre-set maximum deviation, was added to the x, y, and z coordinates of the C^{α} atomic coordinates. This procedure was encapsulated in the program *crystal* . The generation of the backbone atomic coordinates by the MCPB method is dependant on the distance

between two consecutive C^{α} atoms. The normal distance between two consecutive C^{α} atoms is in the range of 3.79 ± 0.04 Å as previously described in chapter two. The larger the random statistical errors introduced into the C^{α} positions, the greater the variation of the distances between two consecutive C^{α} atoms. Therefore random statistical errors introduced into C^{α} coordinates are expected to affect the validity of structures generated from backbone atomic coordinates. With the exception of *cis*-proline, if the distance between any two consecutive C^{α} atoms is larger than 4.1 Å or less than 3.3 Å a complete backbone model structure could not be built by the MCPB method. To overcome this problem which occurs when random statistical errors are introduced to the C^{α} coordinates the distance between any two consecutive C^{α} atoms was limited to within the range 3.7 ± 0.4 Å. and for *cis*-proline to 3.0 ± 0.5 Å . In this way, a complete protein model can be built even when pre-set values of the maximum random statistical error is larger than 2.0 Å.

A number of different sets of C^{α} coordinates were generated with the pre-set maximum deviation gradually increasing by 0.25, 0.5, 1.0 and 2.0 Å. Final models were obtained after energy refinement in which the C^{α} position were not fixed. Up to 300 steps of energy minimisation were carried out. Complete structures were modelled from these 'error-containing' C^{α} co-ordinates and compared with the corresponding X-ray structures of proteins. The r.m.s.d of the model structures with different magnitude of random errors is given in Table 4.8.

Table 4.8. The r.m.s.d (Å) obtained after introducing random errors in the C^{α} coordinates

	Error scale (Å)				
Protein	0.00	0.25	0.50	1.00	2.00
Main chain					
1CRN	0.83	1.24	1.43	1.39	1.38
1ROP	0.55	0.7	0.7	0.84	1.23
1TIM	0.71	1.01	1.05	1.17	1.33
1UTG	0.82	0.91	0.82	0.87	1.28
2CCY	0.56	1.01	1.04	0.96	1.44

2ZTA	0.63	0.56	0.79	0.72	1.28
3FXN	1.16	1.21	1.22	1.38	1.86
4PTI	1.63	1.68	1.56	1.69	1.91
4TNC	1.06	0.98	0.94	0.99	1.48
Overall	0.88	1.03	1.06	1.11	1.51
Side-chain					
1CRN	1.86	1.93	2.21	2.28	2.39
1ROP	2.08	1.99	2.09	2.06	2.35
1TIM	2.3	2.46	2.46	2.72	3.01
1UTG	2.14	2.34	2.02	2.34	2.79
2CCY	2.29	2.49	2.45	2.33	2.84
2ZTA	2.64	2.68	2.43	2.75	2.88
3FXN	2.46	2.7	2.5	2.72	3.41
4PTI	3.14	3.45	3.53	3.37	3.43
4TNC	2.54	2.69	2.51	2.53	3.07
Overall	2.38	2.53	2.47	2.57	2.98
All atoms					
1CRN	1.37	1.48	1.71	1.83	1.92
1ROP	1.51	1.49	1.55	1.57	1.87
1TIM	1.66	1.84	1.85	2.05	2.32
1UTG	1.6	1.75	1.52	1.75	2.15
2CCY	1.59	1.83	1.81	1.71	2.18
2ZTA	1.96	1.98	1.84	2.05	2.27
3FXN	1.9	2.07	1.95	2.14	2.72
4PTI	2.45	2.69	2.71	2.64	2.76
4TNC	1.93	2.01	1.88	1.91	2.39
Overall	1.77	1.90	1.87	1.96	2.39

4.3.5. Comparison of the r.m.s.d. with other methods

In comparing the success of the various methods of modelling proteins from known sequence and C^{α} coordinates with the MCPB/MCPBA method, two criteria have been considered. The first is the accuracy of the model as defined by the r.m.s.d with the X-ray structures and the second the dependence of accuracy on the known atomic positions.

Protein 3FXN has been widely examined by other methods and the r.m.s.d's of the main-chain, side-chain and all atoms of this protein for the different methods are

compared in Table 4.9. The MCPBA method gives a r.m.s.d of the main-chain atoms with a greater error than the other methods but importantly the r.m.s.d of all atoms is better than most of the available methods.

Table 4.9. Comparison of methods for r.m.s.d. (Å) of 3FXN.

Method	Backbone	Side-chain	All atom
Jones and Thirup	0.51	2.41	--a
Reid and Thornton	0.57	--a	1.73
Correa	0.49	--a	1.64
Holm and Sander	0.48	--a	1.57
Mandal and Linthicum	0.46	--a	1.71
Lee and Sander	--a	1.90	--a
Levitt	0.44	1.91	1.37
This work	0.57	2.15	1.55

[a] no data reported.

Holm and Sander[43] have developed a method that combines a data base search for the backbone coordinates and a simulated annealing method for the generation of side-chain coordinates. This method has been tested for eight proteins and the r.m.s.d of the main-chain by this method is between 0.4 Å and 0.6 Å which is comparable with the MCPBA method. The r.m.s.d of the side-chains is 2.21 Å where is comparable to that found in the present study (2.25 Å).

Mathiowetz and Goddard[44] have reported a Dihedral Probability Grid Monte Carlo (DPG-MC) method to build protein models from C^{α} coordinates. The method has been tested for sixty proteins. The r.m.s.d. of backbone atoms is 0.51 Å and 1.73 for all atoms. The method is comparable with the MCPBA method. Another modelling method PROGEN[10] has been used to model sixty proteins and the overall r.m.s.d was 0.53 Å for the main-chain atoms and 1.71 Å for all atoms. These results are not as satisfactory as for the MCPBA method.

4.4. Conclusion

The Monte Carlo Protein Building Anneal method (MCPBA) is a simple method for modelling full atomic structure of proteins starting only with the coordinates of C^{α} atoms and the amino acid sequence. The method is accurate, efficient (40s/residue on an IBM RS/6000) and insensitive to random errors of up to 1 Å in C^{α} coordinates. For main-chain atoms the average r.m.s.d. is 0.45 Å and for all non-hydrogen atoms the r.m.s.d is 1.61 Å. We will subsequently report the use of the MCPBA method to generate the complete atomic coordinates of the coiled-coil structures of rod domain in wool protein. The method is easy to use and does not require a large protein data base as required for homology building. The accuracy of the method is at least comparable with methods previously reported but not as accurate as Levitt's SMM which however requires a large protein data base. The MCPBA method is applicable therefore in situations where no suitable data base is available and complements data base homology modelling.

References

1 Warme P. K., Morgan R. S. (1978) *J. Mol. Biol.* **118,** 273-287. Warme P. K., Morgan R. S. (1978) *J. Mol. Biol.* **118,** 289-304. Claessens M., Cutsem E. V., Lasters I., Wodak S. (1989) *Protein Engineering* **2,** 335-345.

2 Gelin B. R., Karplus M. (1977) *Proc. Nat. Acad. Sci. U.S.A.* **74,** 801-805.

3 McGregor M. J., Islam S. A., Sternberg M. J. E. (1987) *J. Mol. Biol.* **198,** 295-231.

4 Janin J., Wodak S., Levitt M. (1978) *J. Mol. Biol.* **125,** 357-386. James M. N. G., Sielecki A. R. (1983) *J. Mol. Biol.* **163,** 299-361.

5 Ponder J. W., Richards F. M. (1987) *J. Mol. Biol.* **193,** 775-791.

6 The exponential dependence on the size of the protein is a major computational problem. 100 residues with five conformers each results in 7.9×10^{69} conformers.

7 Correa P. E. (1990) *Proteins* **7,** 366-377.

8 Lee C., Subbiah S. (1991) *J. Mol. Biol.* **217,** 373-388.

9 This method does not require a database of known protein structures making its success impressive. A major drawback of the method is that no main-chain atoms are generated.

10 Mandal C., Linthicum, D. S. (1993) *J. Computer-aided Design* **7,** 199-224.

11 Reid L. S., Thornton J. M. (1989) *Proteins* **5,** 170-182.

12 Jones T. A., Thirup S. (1986) *EMBO J.* **5,** 819-822.

13 Blundell T. L., Sibanda B. L., Sternberg M. J. E., Thornton J. M. (1987) *Nature* **326,** 347-352.

14 Levitt M. (1992) *J. Mol. Biol.* **226,** 507-533.

15 Chapter 3

16 Simulated annealing algorithms have been shown to be a valuable tool to overcome the side-chain packing problem. Kirkpatrick S., Gelatt C. D. Jr., Vecchi M. P. (1983) *Science* **220,** 671-680. Brunger A. T. (1988) *J. Mol. Biol.* **203,** 803-816. Subbiah S., Harrison S. C. (1989) *Acta Crystalloger* sect. **A45,** 337-342. Nayeem A. Vila J. Schraga H. A. (1991) *J. Comp. Chem.* **12,** 594-605. Wilson S. R., Cui W., Moskowitz J. W., Schmidt K. E. (1991) *J. Comp. Chem.* **12,** 342-349. Snow M. E. (1992) *J. Comp. Chem.* **13,** 579-584. Press W. H., Flannery B. R., Teukolsky S. A., Vetterling W. T. Combinatorial Minimization: Method of Simulated annealing. In *Numerical Recipes,* Cambridge University Press, (1986) Cambridge. p326.

17 Novotny J., Bruccoleri R., Karplus M. (1984) *J. Mol. Biol.* **177,** 787-818.

18 Dudek M. J., Scheraga H. A. (1990) *J. Comp. Chem.* **11,** 121-151.

19 Momany F. A., Carruthers L. M., Scheraga H. A. (1974) *J. Phy. Chem.* **78,** 1595-1620. Momany F. A., Carruthers L. M., Scheraga H. A. (1974) *J. Phy. Chem.* **78,** 1621-1630.

20 Summers N. L., Carlso W. D., Karplus M. (1987) *J. Mol. Biol.* **196,** 175-198.

21 We have tested the specified ranges of the torsion angles for the particular amino acid residues from a previous statistical analysis of the crystal structures of proteins,

however, the r.m.s.d. of the models are not better than the random selection within the ranges of -180° to 180°.

22 Zimmerman S. S., Pottle M. S., Nemethy G., Scheraga H. A. (1977) Macromolecules **10,** 1-9.

23 Momany F. A., McGuire R. F., Burgess A.W., Scheraga H. A. (1975) *J. Phy. Chem.* **79,** 2361-2380. MacArthur M. W., Thornton J. M. (1991) *J. Mol. Biol.* **218,** 397-412.

24 $E_{vdw} = \varepsilon_0[(r_0/ r_{ij})^{12} - 2(r_0 / r_{ij})^6]$ where r_{ij} is the distance of a given pair of atoms i and j, ε_0 and r_0 are constant parameters describing respectively, the depth of the energy well, and the equilibrium inter atomic distance for van der Waals interaction of a given pair of atoms (the parameters ε_0 and r_0 are taken from OPLS force field developed by Jorgensen (Jorgensen W. L., Rives T. J. (1988) *J. Am. Chem. Soc.* **110,** 1657-1666). The interaction becomes infinite as r tends to zero. (Atkins. P. W. (1986) *Physical Chemistry*, p587, Freeman, New York). Such infinite energy barriers would block simulated annealing from exploring the conformational solution space, and have therefore been truncated to a maximum value of 7 kcal/mol (1cal = 4.184 J) for each pairwise interaction, (similar to the 'soft atoms' described by Levitt M. (1983) *J. Mol. Biol.* **170,** 723-764).

25 Nemethy G., Pottle M. S., Scheraga H. A. (1983) *J. Phys. Chem.* **87,** 1883-1887.

26 Weiner S. J., Kollman P.A., Case D. A., Singh U. C., Ghio C., Alagona G., Profeta S., Weiner P. (1984) *J. Am. Chem. Soc.* **106,** 765-784.

27 Jorgensen W. L., Rives T. J. (1988) *J. Am. Chem. Soc.* **110,** 1657-1666.

28 The energy parameters can be chosen in the crystal.com file from three force fields, ECEPP/2, AMBER and OPLS. We have tested the energy parameters of these three force fields on protein 1CRN, 1ROP, 3FXN and 2ZTA respectively. The result has shown the OPLS force field is the best for our purpose.

29 Chapter 3

30 This sort of distortion was not observed when we averaged the atomic co-ordinates of the backbone. The same problem was observed by Levitt.[14]

31 Macromodel software makes use of a standard inter atomic potential energy function, consisting of bond stretching, angle bending, tosional, non-bonded interactions and electrostatic interactions. The OPLS/AMBER non-hydrogen force field and GB/SA water model are selected for the energy minimisation of the mean structures. The default values of the cut off distances of each interaction atomic pair in Macromodel/Batchmin is used.The conjugate gradient method is selected for energy minimisation. The method is commonly used for minimisation of proteins and macromolecules. During energy minimisation, no hydrogen atoms are included and the C^{α} co-ordinates in the mean structure are fixed in their position defined by the X-ray structure. Some 200 steps of energy minimisation are generally required to correct the angles and bond lengths of the mean structure. A default value of 300 steps is used to ensure the structures are corrected.

32 Still W. C., Tempczyk A., Hawley R. C., Hendrickson T. (1990) *J. Am. Chem. Soc.* **112,** 6127-6129.

33 This is done using Macromodel/BatchMin .com file by adding the command FXAT after command READ and setting zeros in the x, y, and z columns.

34 Bernstein F. C., Koetzle T. F., Williams E. J. B., Meyer Jr. E. F., Brice M. D., Rodgers J. R., Kennard O., Shimanouchi T., Tasumi M. (1977) *J. Mol. Biol.* **112,** 535-542.

35 Lesk A. M. Protein Architecture, (1991) *Oxford University Press* p15.

36 O'Shea E. K., Rutkowski R., Kim P. S. (1989) *Science* **243,** 538-542.

37 Banner D W., Michael K., Tsernoglou D. (1987) *J. Mol. Biol.* **196,** 657-675.

38 A value of the r.m.s.d of the main-chain atoms of 0.43 Å was previously reported in chapter three but in that instance the mean backbone structrure with the C^{α} carbons fixed was minimised without the side chains.

39 Engh R. A., Huber R. (1991) *Acta Cryst.* **A47**, 392-400.

40 Pascual-Ahuir J. L., Silla E (1990) *J. Comp. Chem.* **11,** 1048-1060. Silla E., Tunon I., Pascual-Ahuir J. L. (1991) *J. Comp. Chem.* **12,** 1077-1088. Floris F. M., Tomasi J., Pascual-Ahuir J. L. (1991) *J. Comp. Chem.* **12,** 784-791.

41 The SAS area is more closely related to side-chain conformation of the proteins than VDW surface area. The SAS area with the probe of radius 1.4 Å (water) is sensitive to the side-chain conformation.

42 Chothia C. (1975) *Nature* **254,** 304-308.

43 Holm L., Sander C. (1991) *J. Mol. Biol.* **218,** 183-194.

44 Mathiowetz A. M., Goddard W. A. (1995) *Protein Science* **4,** 1217-1232.

Chapter Five

Modelling studies of the coiled-coil protein in wool

Summary

A full atomic model of the rod domain of wool has been established from the amino acid sequences of proteins 7c and 8c-1 using the MCPB/MCPBA method. For the particular knob-hole heptad repeat model investigated in this study the single α-helical chain the rise per residue is 1.464 Å, the twist angle per residue 102.999°, the number of residues per turn is 3.524 and the pitch 5.171 Å. For the four coiled-coil helical segments of the rod domain the pitch is in the range 124 Å to 192 Å and the radius varies between 5.24 Å to 5.92 Å. The inter-chain interaction energy is evaluated for van der Waals non-bonded, electrostatic and hydrogen bonding interactions. The optimum relationship of the α-helical chains to each other established the heptad repeat interaction; 34% of the leucine residues are located at the *d* position. Of the potential backbone hydrogen bonds in the α-helix between residues four apart 18% have a distance between a donor NH nitrogen and acceptor carbonyl oxygen greater than 3.5 Å. The hydrogen bonds between the side-chains of the two α-helices in the model are largely between Arg and Glu, Arg and Asp and Glu and Asp. The distances between the C^{β} atoms of cysteine residues are > 4.5 Å and outside that required for formation of disulfide bonds. The interchain interaction of charged residues with apolar, polar and charged residues in the *a-a, a-d, d-d,* and *d-a* heptad positions accounts for 70% of the interaction energy. The solvent accessible surface areas and volumes for the rod domain and for each of the coiled-coil helical segments are reported.

5.1. Introduction

The commercial importance of wool has provided much of the impetus for undertaking research into keratin intermediate filament (IF) structures over the past fifty years. However, the structural problems posed by wool have proved to be extraordinarily complex. Many of the recent advances in keratin structure have arisen from investigations into non-keratin members of the IF family of proteins. These non-keratin proteins have given a wealth of complementary data and have permitted new insights into the structure of keratin proteins. In wool protein, a filament-matrix composite is characterised by a high content of covalent disulfide bonds that weld the constituent rod domains and matrix into a mechanically stable structure.[1] The extent of disulfide bonding has made chemical investigation difficult.

Many features of keratin IF structure have now gained wide acceptance.[2,3,4,5] A notable feature of the coiled-coil rod domain in wool is the repeat heptad of residues. Coiled-coil structures are found not only in fibrous proteins e.g. keratin[6] and tropomyosin[7] but also in globular proteins.[8] The coiled-coil helical structure in the rod domain of wool is not continuous and is interrupted by several short sections of non-helical structure (Figure 5.1).[9] The N- and C-termini regions are non-helical. Chemical and biochemical data have shown unequivocally that the coiled-coil unit of keratin and other IF proteins consists of two chains[10] rather than three chains as previously suggested.[11] A typical rod domain consists of two different α-helical chains, 8c-1 (Type I - acidic) and 7c (Type II - basic) that are parallel and defined in a direction from the N- to the C-termini in axial register.[12,13] The rod domain of wool keratin is ca. 470 Å in length and consists of two segments (segment 1 and 2) each about 220 Å in length. Segment 1 comprises two heptad-containing segments (1A and 1B) separated by a variable non α-helical link, L1. Likewise segment 2 contains a pair of heptad-containing segments (2A and 2B) separated by a length of non α-helical link, L2. Segments 1 and 2, in turn, are connected by a non-helical link, L12.

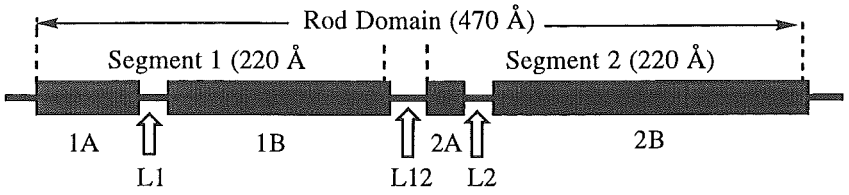


Figure 5.1 (a). The structure of the coiled-coil rod domain in wool protein. (b) the chains of the rod domain

The coiled-coil structure is stabilised by the interactions between the intertwining helices. Optimal packing is achieved by fitting the 'knobs' of one helix into the 'holes' of the other.[14] The presence of a periodic disposition of apolar residues at positions *a* and *d* within a heptad repeat of the form, $(a, b, c, d, e, f, g)_n$ signals the potential for the interlocking of side-chains between the helices in the coiled-coil. This type of periodicity is common in fibrous proteins and is characteristic of most membrane proteins as well as in a wide range of globular proteins. The coiling of two helical chains about each other optimises inter-chain packing of apolar residues within the coiled-coil. It is these interactions which define direction and the relative chain alignment between the various coiled-coil ropes that make up the wool fibre.

The simplicity and elegance of the coiled-coil motif has spawned a number of modelling studies. The mathematical description of an ideal coiled-coil model was first developed by Crick[14] and his equations form the basis of most studies. Regular mathematically defined coiled-coils have been subjected to energy minimisation. Parry and Suzuki[15] built a model using Crick's equations of two stranded coils of poly-L-alanine with a pitch of 186 Å and examined a range of radii of the coiled-coil. A minimum energy was found for a radius of 4.0 Å. McLachlan[16] proposed a model of the murein lipoprotein of *Escherichia coli*[17] as a two stranded coiled-coil with 58 residues in each chain and a pitch of 186 Å. Energy calculations for a series of coiled-coils with different initial radii defined the lowest energy structure with a radius of the coiled-coil of 4.125 Å.

A detailed coiled-coil structure for specific protein sequences is required to address questions concerning the structural basis of stability and sequence specificity for coiled-

coil interactions.[18] To date only a few data sets of the X-ray structures of coiled-coil proteins are on deposit in the PDB.[19] Structures are available for Tropomyosin filaments,[20] GCN4 (Leucine Zipper)[21] and 1ROP.[22] X-ray derived coordinates for the rod domain of wool protein have not been reported. For the few existing crystal structures the pitch of the coiled-coils vary in a range 137 Å to 173 Å and the radii of the coiled coil between 4.0 to 4.8 Å. The X-ray structure of the protein tropomyosin filament,[23] which is a two stranded parallel coiled-coil of 400 Å in length, has the pitch 137 Å and a radius of 4.05 Å.[20] The protein yeast transcriptional activator GCN4 (2ZTA),[21] a short coiled-coil with two parallel chains containing 33 residues in each chain, has a pitch of 144 Å and a radius of 4.8 Å. The protein 1ROP, a monomer with 56 residues with an anti-parallel coiled-coil structure, has a pitch of 172.5 Å and radius of 4.6 Å.[22]

Our efforts have been directed to creating a protein modelling algorithm to generate the coiled-coil structure of wool protein in a study directed to explaining details of the structure including the interaction energy between the two strands of the coiled-coil. The mathematical equations for an idealised coiled-coil derived by Crick[14] were first used to generate all the atomic coordinates of the backbone atoms of a coiled-coil. These equations have been used by Parry[15] to calculate the dihedral angles ϕ and ψ of the backbone helical chains in the coiled coil. These values are the similar for each amino acid namely -52° ± 1° and -51° ± 1° respectively. However our investigation of the α-helical motif region in the native crystal structures of Tropomyosin filaments, GCN4 (Leucine Zipper) and ROP show the dihedral angles of the protein backbones have a distribution of values in the range -60° ± 30° and -40° ± 20° respectively.[24] For this reason, we have chosen to generate only the C^{α} coordinates using Crick's equations and the coordinates of the remaining atoms of the backbone are generated using a Monte Carlo (MCPB) method. In this way, the dihedral angles of the backbone show deviations in values similar to those observed in crystal structures.

5.2. Computational methods and strategy

5.2.1. Generation of atomic coordinates

Our strategy for the generation of the model of the coiled-coil helical rod domain in wool protein has been to obtain the C^{α} atoms of coiled-coil model using Crick's equations[25] which are encapsulated in the program *carb*. The program reads the amino acid sequences[26,27,28] of 7c and 8c-1 (Table **5.1**) from the sequence files (seq1.dat and seq2.dat) and the parameters (pitch, radius, phase angles etc.) and the lengths of linking segments (L1, L12 and L2) between two helical segments are defined in the carb.com file.

Table 5.1. The residue number in the segments of the rod domain of wool

Segment	Sequence in wool protein	Sequence in model	Chain 8c-1 Number of residues	Sequence in wool protein	Sequence in model	Chain 7c Number of residues
1A	56 - 90	1 - 35	35	110 - 144	1 - 35	35
L1	91 - 101	36 - 46	11	145 - 154	36 - 45	10
1B	102 - 202	47 - 147	101	155 - 255	46 - 146	101
L12	203 - 218	148 - 163	16	256 - 272	147 - 163	17
2A	219 - 237	164 - 182	19	273 - 291	164 - 182	19
L2	238 - 245	183 - 190	8	292 - 299	183 - 190	8
2B	246 - 366	191 - 311	121	300 - 420	191 - 311	121

5.2.1.1. Generation of C^{α} coordinates of the helical segments

Both chains in the rod domain of wool including the residues in the random segments (L1, L2 and L12) contain 622 residues. The residues from 1 to 55 of 8c-1 and 1 to 109 of 7c and the residues after residue 366 of 8c-1 and after 420 of 7c are not in the rod domain. The coiled-coil chain of 8c-1 is from 56Lys to 366Leu and for 7c is from 110Lys to 420Leu.[26-28] In our model the first amino acid residue in the coiled-coil structure, namely 110K of chain 7c, is defined as residue 1. The generation of the C^{α} coordinates of the rod domain of wool was therefore commenced from residue 1 to 311 for 7c and similarly for 8c-1.

The C^{α} coordinates of the helical segments were generated in the following order: 1A (7c) / 1A (8c-1), 1B (7c) / 1B (8c-1), 2A (7c) / 2A (8c-1) and 2B (7c) / 2B (8c-1).2)

Crick's[14] values of pitch and radius for keratin, namely 186 Å and 5.5 Å, were selected in the first instance. Fraser has suggested that the radii of the coiled-coil rod domain in wool protein is between 5.2 and 5.5 Å.[6] We varied the values of pitch and radii. A value of pitch of 200 Å [29] and radii in the range of 4.8 to 5.5 Å were found to be most satisfactory as starting values for generating the model of the rod domain and these values change on energy minimisation.

5.2.1.2. Generation of C^{α} coordinates of the linking segments

After the C^{α} coordinates of the coiled-coil helical segments (1A, 1B, 2A and 2B) have been established, the random coil regions between two coiled-coil helices are generated (L1, L12 to L2). The linking segments are considered to be in a random coil motif.[6] To meet the requirements for the length of the rod domain (470 Å), the lengths of the segments L1, L2 and L12 are first set proportional to the number of residues in each linkage region, namely to values of 15, 30 and 15 Å and defined as such in the .com file. The first C^{α} position of a linking segment is built from the last three C^{α} atoms in the preceding coiled-coil segment.[30] The distance between C^{α} atoms is randomly selected from a value between 3.7 and 3.9 Å, and the C^{α}_{i-1},C^{α}_{i},C^{α}_{i+1} angle from the range 90° to 140° and the torsional angle of four consecutive C^{α} atoms from the range -180° to 180°.

After the generation of the last C^{α} atom of the joining segment, four criteria are used to determine whether the coordinates of the linking region are accepted or discarded. The first criterion is the distance between the last C^{α} atom of the joining segment and the first C^{α} atom of the second helical segment. For the coordinates to be accepted, the distance between two consecutive C^{α} atoms must be in the range 3.7 - 3.9 Å, the C^{α}_{i-1},C^{α}_{i},C^{α}_{i+1} angle and the angle of the last C^{α} atom of the joining segment and the first consecutive two C^{α} atoms of the second helical segment must each be in the range 90 - 140°. The final criterion is that the distance between any two C^{α} atoms in the joining segments must be at least 3.0 Å apart to avoid the 'crash' of any two amino acid residues when the remaining atoms of the backbone or the side-chain are attached. If any of these

criteria are not met the joining segment is regenerated. Output of the calculated structure is in Macromodel format.

5.2.1.3. Generation of the backbone atoms

The MCPB method[31] developed to generate the remaining atoms of the backbone from the C^{α} coordinates (see Chapter 3) was used to generate ten independent backbone models of the rod domain using different random seeds. These structures were averaged to give a mean backbone model of the coiled-coil which was not subject to energy minimisation at this stage.

5.2.1.4. Generation of the side-chain conformations

The MCPBA[32] simulated annealing technique was used to generate the side-chain conformations for the coiled-coil and the linking regions. Only vdW non-bonding energy interactions using the OPLS force field parameters[33] are included in assessing energy during simulated annealing. For the generation of the side-chain ensemble of conformations the initial temperature for simulated annealing was set in the first instance at 1,500°C. Simulated annealing stops when one of three conditions are met; a maximum number of conformations are generated (20000), the ratio of the number of accepted conformations to the number of conformations generated is less than 0.2 or the number of conformations consecutively rejected (1000) is met. In general, about four thousand conformations are generated, about five hundred conformations are accepted and three hundred of them collected in a .dat file. The collected conformations are averaged to generate a mean structure.

5.2.2. Energy minimisation.

Energy minimisation for each of the mean models with varying radii is carried out using Macromodel/Batchmin5K[34] and the OPLS/Amber force field[33] and GB/SA water model.[35] The cut-off default distances for vdW and electrostatic interactions is 6.0 Å and 12.0 Å respectively. The conjugate gradient method is used for energy minimisation[36] which is carried out without constraints on any of the atoms. The total number of non-

hydrogen atoms in the coiled-coil and linking regions of the rod domain is 5061 which exceeds the maximum number of non hydrogen atoms (5000) allowed for energy minimisation in Macromodel/Batchmin5K. To ensure the coiled-coil segments of the rod domain can be minimised the side-chains (excluding C^{β} atoms) of selected residues in the random segments; 40E in L1, 152R in L12, 184R, 185R, 190W in L2 in the 8c-1, 38R, 42E in L1, 156K, 161R in L12, 184R, 190W in L2 in the chain 7c are removed and replaced by C^{β}. This reduces the total number of non hydrogen atoms including capped groups (-NCH_3) at C-terminal of each chain in the rod domain to 4994. These residues are expected to have little influence on the pitch or radius of the coiled-coil in the rod domain.

The large number of residues in the coiled-coil rod domain still generates a problem during energy minimisation of the mean structures of the complete rod domain which can not be directly performed even though the number of the atoms (4994) is less than 5000. The mean structures can have poor stereochemistry with some 30% of the bonds being shorter than normal resulting in the energy of the structure being so large as to overflow capacity of the Macromodel/Batchmin5K. To generate the final model of the wool protein four steps are carried out. The *first* step in the energy minimisation is directed to overcome the above problem. The structure is cut into seven pieces; 1A, L1, 1B, L12, 2A, L2 and 2B respectively by using the program *divide* and fifty iterations of energy minimisation for each piece are carried out to correct bond lengths and angles. The segments are rejoined by using the program *combine*. The *second* step, namely energy minimisation of the rod domain, is then carried out. After energy minimisation of each of the mean structures it was found that for some regions of the coiled-coil segments the heptad repeat was lost. The strategy developed to overcome this problem was to select and link segments from the different structures generated from models with differing initial radii which had retained the heptad repeat to produce a final model. To do this it was found necessary to again cut each of the minimised models of the rod domain into the four coiled-coil helical (1A, 1B, 2A and 2B) and three link segments (L1, L12 and L2). In the *third* step the four coiled-coil helical segments of the different models are separately minimised and the structures examined to find the segments where the heptad

repeats remain intact. The *fourth* step involves joining the segments of lowest conformational energy where the heptad repeat is intact. The final structure, formed in this way, was subjected to energy minimisation and in all instances studied this occurred without loss of the heptad repeats.

5.2.3. Computer programs

The helical parameters of the minor and major helices are calculated using the program *axisc* based on the algorithm developed by Kehn,[37] and written in FORTRAN to read a file in Macromodel format (Appendix 2). The distribution of the charged amino acid residues for the rod domain of wool protein have been previously investigated by Parry and Fraser.[5] The distribution of the charged residues on Type I chain 8c-1 and Type II chain 7c are separately calculated for the model using the program *mtranscoil.* The distribution of the dihedral angles in the model coiled-coil structure are calculated using the program *tors.* Over 88% of the residues in the rod domain of wool are considered to be in an α-helical motif.[38] There are three main factors which determine conformation in fibrous proteins; the interaction between charged, apolar, polar and aromatic side-chains; hydrogen bonding interactions and disulfide bonding. These interactions were analysed using the program *mtranscoil.* The solvent accessible surfaces and volumes of the molecular models are calculated for the rod domain and for each of the four helical segments using the program *Gepol.*[39]

5.3. Results and discussion

5.3.1. Determination of the heptad repeat of the coiled-coil helices

The definition of the heptad repeat in coiled-coil proteins is important in defining inter chain interactions of the interlocking residues.[40] The *a* and *d* positions of a heptad repeat are frequently occupied by apolar residues, especially leucines.[40] For example, in both helical chains of the coiled-coil protein 2ZTA, the percentage of leucine residues at the *d* position is 67% (8/12). For Tropomyosin[41] 34% of the leucine residues are in the *d* position. The percentage of leucine residues occupying the *a or d* position in a heptad repeat can be used as a criterion to judge if the heptad is part of a coiled-coil structure.

We located the knob-hole heptad repeats in the coiled-coil rod domain in the following way. The phase angles ϕ_{11} and ϕ_{12} of the C^{α} atoms of the α-helical chains respectively were rotated through 360° with corresponding longitudinal movement of each C^{α} atom. The knob-hole structures were investigated by analysis if the distance matrixes. The structure with the greater percentage of leucine residues occupying positions *a* and *d* was modelled and the structure examined by energy calculations (Figure 5.2).

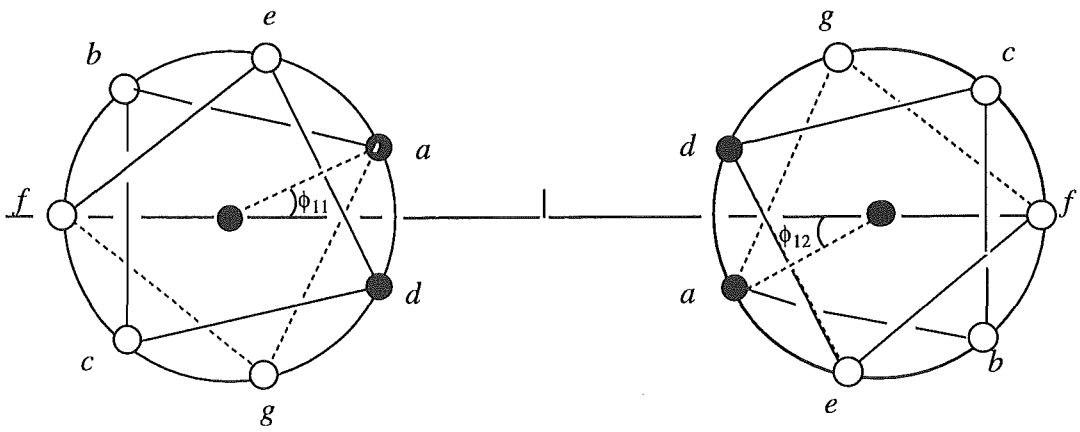


Figure 5.2 The heptad repeats of a coiled-coil

A distance matrix is established between pairs of seven C^{α} atoms in each α-helical chain for each relative orientation of the two chains. From this multitude of matrices, the residues represented by C^{α} atoms located at the inter-face of the coiled-coil will have the shortest distances when located at the *a* and *d* positions in a knob-hole structure. The programme '*carb* ' is used to analyse the matrices. If two closest residues in the same chain are separated by two positions the first residue is in the *a* position and the second residue in *d*. If the residues in the same chain are separated with three amino acid residues, the first residue is located in the *d* position and the second residue in the *a* position. If the two residues are not separated by 2 or 3 amino acid residues then they do not represent a heptad repeat.

5.3.1.1. Location of the heptad repeats of 2ZTA.

As an example, a distance matrix of the C^{α} atoms for the first heptad repeat in the X-ray structure of the protein 2ZTA is shown in the Table 5.2. The residues of the first chain are listed in the first row and the residues of the second chain in the first column.

Table 5.2. The distance matrix (Å) of C^{α}'s in the first heptad repeat of 2ZTA

Residue	1 R	2 M	3 K	4 Q	5 L	6 E	7 D
1 R	10.164	7.762	10.994	12.564	9.921	10.045	13.732
2 M	7.562	6.313	9.860	10.444	7.826	9.375	12.628
3 K	10.685	9.773	13.072	13.020	9.833	11.559	14.671
4 Q	12.631	10.554	13.322	13.597	9.979	10.556	13.850
5 L	10.139	7.963	10.117	10.075	6.339	6.924	10.054
6 E	10.338	9.527	11.744	10.530	6.956	9.140	11.411
7 D	13.979	12.744	14.890	13.892	10.138	11.471	13.778

For the first seven amino acids of each chain, which have the same sequence, the shortest distance in the matrix is 6.313 Å between the second residue (Met) of one helix and the second residue (Met) in the other helix. The second shortest distance is 6.339 Å between the fifth residue (Leu) of one chain and the fifth residue (Leu) of the other chain. The second residue (Met) in the first chain has two residues between it and Leu so Met and Leu are in the *a* and *d* positions. In the second chain, the same situation occurs with 2Met and 5Leu which are located at the *a* and *d* positions respectively. After the *a* and *d* positions in the first heptad repeat in each chain are defined the other residues in all positions for the coiled-coil follow. For 2ZTA with parallel identical chains, the residues in the *a* and *d* positions in the two helices are: 2M(*a*), 5L(*d*), 9V(*a*), 12L(*d*), 16N(*a*), 19L(*d*), 23V(*a*), 26L(*d*), 30V(*a*).[42] We subsequently generated some 250 different models of this protein by rotating and 250 by screwing both chains. The distance matrices obtained by rotating the two chains in each of these models with respect to each other were examined using the program *carb*. For the models obtained by a single rotation, the heptad repeat located by the method outlined above is the same as found in the X-ray structure.

5.3.1.2. Location of the heptad repeats in the rod domain

The various possible heptad repeats of the residues in the rod domain are similarly found by rotating the two phase angles of the C^{α} atoms in each chain as indicated above. The first chain is fixed and the second chain rotated in increments of 10°. After each rotation a matrix for the C^{α}/C^{α} interchain distances is determined. The first chain is subsequently rotated by increments of 10° and for each rotamer the second chain is rotated in increments through 360°. This process is carried out for each of the four coiled-coil helical segments (1A, 1B, 2A and 2B) of the rod domain of wool protein. The matrix with the highest percentage of leucine residues occupying the *a* and *d* positions in each of the coiled-coil helical segments for the two chains is determined. For this structure, which is only one of 49 possibilities the phase angles which result in the most satisfactory interaction of heptad repeats are 150° (8c-1) and 150° (7c) for segment 1A, 250° (8c-1) and 200° (7c) for segment 1B, 200° (8c-1), and 200° (7c) for segment 2A, and 150° (8c-1) and 250° (7c) for segment 2B respectively. Various other structures will be generated for subsequent analysis, but are not included in this thesis.

The percentage of leucine residues occupying the positions *a* and *d* in each of coiled-coil helical segments for the above structure are given for the four helical segments 1A, 1B, 2A and 2B respectively in Table 5.3. The total number of leucine residues in the

Table 5.3. The percentage of leucine in positions *a* and *d* of the heptad repeats.

	7c		8c-1	
Helical Segment	Position *a*	Position *d*	Position *a*	Position *d*
1A	1/5	3/5	1/6	3/6
1B	3/12	6/12	7/16	1/16
2A	0/0	0/0	1/3	2/3
2B	1/15	5/15	5/15	6/18
Overall	5/32	14/32	14/43	11/43

coiled-coil helicals egments of the rod domain in the two chains is 75 (32 leucines in chain 7c and 43 leucines in chain 8c-1). In chain 7c, the percentage of leucines at position

a is 15.6% and at *d* is 44% and in chain 8c-1 the percentages of leucine in positions *a* and *d* are 33% and 28% respectively. There is no leucine residue in the helical segment 2A of chain 7c. The overall percentage of leucines occupying positions *a* and *d* in the two chains, 7c and 8c-1, is 25.3% and 34.7% respectively.

The sequences of the amino acid residues in the heptad repeats for this particular heptad repeat structure of the coiled-coil rod domain in wool protein are given in Table 5.4. A heptad repeat of the residues of a coiled-coil segment in chain 8c-1 was first published by Dowling et al.[43] and derived from the amino acid sequences of the chain 8c-1 by preferential assignment of the non-polar residues to positions *a* and *d*. Our studies of the heptad repeats of residues in the helical segments 1A, 2A and 2B of the chain 8c-1 are consistent with Dowling's[43] report. However for segment 1B of chain 8c-1, the heptad repeats in this structrue are different with the position *a* in our model being position *d* in Dowling's sequence. In Dowling's heptad repeat, the percentage of leucine in the *a* and *d* position in the segment 1B of chain 8c-1 is 18% and 44% respectively while in our model the percentage of leucine in the *a* and *d* positions is 44% and 6% respectively. However for our model of segment 1B of chain 7c, which Dowling[43] did not take into consideration, 50% of the leucines are located at position *d* and 25% of leucines are in the position *a*. A 3D diagram and contour map showing the relative rotation of the axis of the two minor helices with the percentage of the leucine residues located at the *a* and *d* positions in the segment 1B is reported in Figure 5.3. In subsequent study the Dowling structure will be further investigated.

Table 5.4. Heptad repeats in the coiled-coil model

Chain 8c-1							Chain 7c						
a	*b*	*c*	*d*	*e*	*f*	*g*	*a*	*b*	*c*	*d*	*e*	*f*	*g*
		1	A						1	A			
				K	E	T					K	E	Q
M	Q	F	L	N	D	R	I	K	C	L	N	N	R
L	A	S	Y	L	E	K	F	A	A	F	I	D	K
V	R	Q	L	E	R	E	V	R	F	L	E	Q	E
N	A	E	L	E	S	R	N	K	L	L	E	T	K
I	L	E	R				L	Q	F	F			
		1	B						1	B			
				Y	Q	S	E	P	L	F	E	G	Y
Y	F	R	T	I	E	E	I	E	T	L	R	R	E
L	Q	Q	K	I	L	C	A	E	C	V	E	A	D
A	K	S	E	N	A	R	S	G	R	L	S	S	E
L	V	V	Q	I	D	N	L	N	H	V	Q	E	V
A	K	L	A	A	D	D	L	E	G	Y	K	K	K
F	R	T	K	Y	E	T	Y	E	Q	E	V	A	L
E	L	G	L	R	Q	L	R	A	T	A	E	N	E
V	E	S	D	I	D	G	F	V	A	L	K	K	D
L	R	R	I	L	D	E	V	D	C	A	Y	V	R
L	T	L	C	K	S	D	K	S	D	L	E	A	N
L	E	A	Q	V	E	S	S	E	A	L	I	Q	E
L	K	E	E	L	I	C	I	D	F	L	R	R	L
L	K	S	N	H	E	E	Y	Q	E	E	I	R	V
E	V	N	T	L	R	S	L	Q	A				
		2	A						2	A			
		D	L	N	R	V			N	M	D	C	I
L	N	E	T	R	A	Q	V	A	E	I	K	A	Q
Y	E	A	L	V	E	T	Y	D	D	I	A	S	R
		2	B						2	B			
			Y	I	R	Q			Y	R	S	K	C
T	E	E	L	N	K	Q	E	E	I	K	A	T	V
V	V	S	S	S	E	Q	I	R	H	G	E	T	L
L	Q	S	C	Q	T	E	R	R	T	K	E	E	I
I	I	E	L	R	R	T	N	E	L	N	R	V	I
V	N	A	L	Q	V	E	Q	R	L	T	A	E	V
L	Q	A	Q	H	N	L	E	N	A	K	C	Q	N
R	D	S	L	E	N	T	S	K	L	E	A	A	V
L	T	E	T	E	A	R	T	Q	A	E	Q	Q	G
Y	S	C	Q	L	N	Q	E	V	A	L	N	D	A
V	Q	S	L	I	S	N	R	C	K	L	A	G	L
V	E	S	Q	L	A	E	E	E	A	L	Q	K	A
I	R	G	D	L	E	R	K	Q	D	M	A	C	L
Q	N	Q	E	Y	Q	V	L	K	E	Y	Q	E	V
L	L	D	V	R	A	R	M	N	S	K	L	G	L
L	E	C	E	I	N	T	D	I	E	I	A	T	Y
Y	R	G	L	L	D	S	R	R	L	L	E	G	E
E	D	C	K	L			E	Q	R	L			

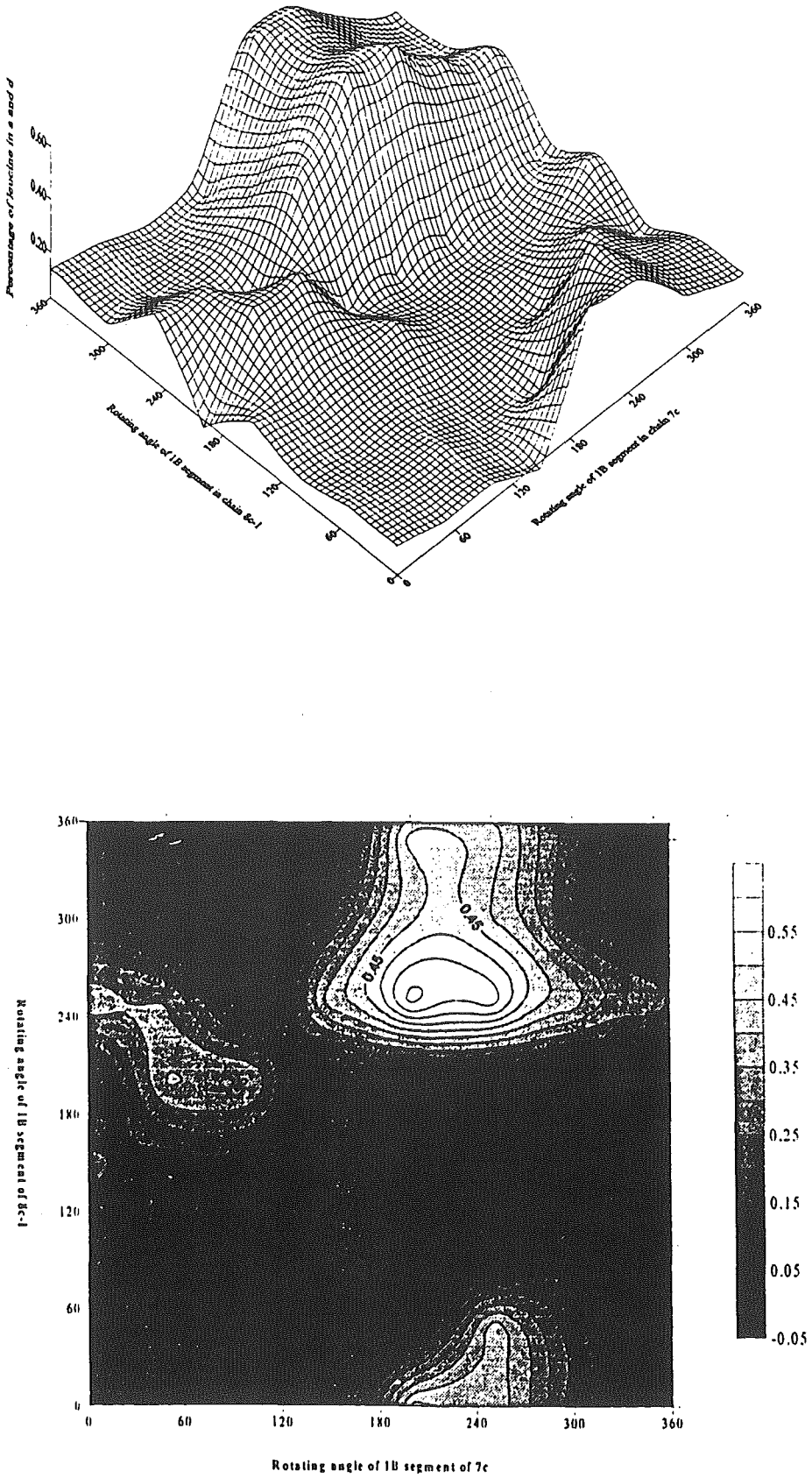


Figure 5.3. A 3D diagram and contour map showing the relative rotation of the axis of the two minor helices as a function of the percentage of the leucine residues located at the *a* and *d* position in the segment 1B.

The highest percentage of leucine in the *a* and *d* positions of the coiled-coil segments 1B of both the 8c-1 and 7c chains is 61%. The structure is defined by the rotation angle for chain 7c of 200° and for 8c-1 of 250°. This structure contains 36% of the leucines in the *a* position and 25% in the *d* position in both chains. There are 276 residues in each of the parallel chains in the coiled-coil helical segments of the rod domain. The distribution of amino acid residues at the seven positions of the heptad is shown in Table 5.5.

Table 5.5. The percentage of amino acid residues in the seven positions of the heptad.

Amino	Num*	*a*	*b*	*c*	*d*	*c*	*f*	*g*
Ala	43	0.093	0.093	0.279	0.093	0.186	0.209	0.047
Arg	45	0.111	0.222	0.089	0.044	0.156	0.200	0.178
Asn	30	0.100	0.233	0.033	0.100	0.200	0.200	0.133
Asp	26	0.077	0.192	0.154	0.115	0.000	0.308	0.154
Cys	15	0.000	0.067	0.400	0.133	0.133	0.067	0.200
Gln	42	0.048	0.262	0.095	0.119	0.143	0.190	0.143
Glu	81	0.099	0.185	0.123	0.099	0.160	0.173	0.160
Gly	12	0.000	0.083	0.333	0.083	0.000	0.333	0.167
His	4	0.000	0.000	0.500	0.000	0.500	0.000	0.000
Ile	28	0.250	0.071	0.107	0.071	0.357	0.071	0.071
Leu	75	0.253	0.040	0.107	0.347	0.133	0.013	0.107
Lys	33	0.061	0.242	0.030	0.242	0.152	0.152	0.121
Met	4	0.500	0.000	0.250	0.250	0.000	0.000	0.000
Phe	11	0.273	0.091	0.364	0.273	0.000	0.000	0.000
Pro	1	0.000	1.000	0.000	0.000	0.000	0.000	0.000
Ser	28	0.107	0.071	0.357	0.036	0.143	0.143	0.143
Thr	24	0.083	0.083	0.167	0.208	0.000	0.208	0.250
Tyr	18	0.333	0.000	0.056	0.222	0.222	0.000	0.167
Val	32	0.250	0.156	0.031	0.094	0.094	0.094	0.281

Alanine prefers position *c* (27.9%) and *f* (20.9%). Arg prefers position *b* 22.2% and *g* 20.0%. Asn and the Asp residues favour positions *b* and *f*. Cys is concentrated in position *c* (40%) and *f* (20%). Gln is favoured at *b* (26.2%) and *g* (19.0%). Glu is the most common residue (81) and is found in positions *b, c, e, f, g* with a frequency

between 10% to 20%. Gly is not found in positions *a* or *e*. A few glycine residues are in position *d* (8.3%), but it most commonly occurs at positions *c* (33.3%) and *f* (33.3%). There are only four histidine residues in the helical segments of wool; two occupy a *c* position and two a *e* position. Ile is favoured at positions *a* (25%) and *e* (35.7%). The percentage of leucine residues in position *d* is 34.7% and in position *a* is 25.3%. Lys residues prefers the *b* and *d* positions, both have a occupancy of 24.2%. Met prefers positions *a, c* and *d*. There is only one proline residue and it occupies position *b*. Ser and Cys residues are preferentially located at position *c* with the same percentage occupancy (35.7%). Tyr is in positions *a* (33.3%), *g* (25.0%), *d* (20.8%) and *f* (20.8%). Valine residues are in positions *a* (25.0%), *g* (28.1%) and *d* (9.4%).

The percentage of non-polar (Ala, Gly, Ile, Leu, Met, Phe, Pro and Val), polar (Asn, Cys, Gln, His, Ser, Thr, and Tyr) and charged residues (Arg, Glu, Asp, and Lys) located in each of the heptad positions have been determined. There are 76 residues in *a* positions, 43 (56.6%) are non-polar residues, 16 (21.1%) are polar residues and 17 (22.3%) are charged residues. In the *b* position, 21.8% of the residues are non-polar residues, 29.5% polar residues and 48.7% charged residues. In the *c* position, the percentage of non-polar, polar and charged residues are 41.3%, 35.0 and 23.7% respectively. In the *d* position, the percentage of non-polar, polar and charged residues are 49.4%, 24.7 and 25.9% respectively. In the *e* position, the percentage of non-polar, polar and charged residues are 38.8%, 30.0 and 31.2% respectively, in the *f* position 24.1%, 30.4 and 45.5% respectively and in the *g* position 29.5%, 33.3 and 37.2% respectively. The non-polar residues prefer the *a* and *d* positions, the charged residues the *b, f, g* positions and the polar residues the *c, e, f, g* positions.

5.3.2. Refinement of the models

The series of models of the rod domain with different initial radii (4.8, 5.0, 5.2 and 5.5 Å), and a pitch of 200 Å show that conformational energy is not particularly sensitive to the radius of the coiled coil (Table 5.6).

Table 5.6. The energy of the models.

Model	Radius (Å)	N_1	N_2	E kcal/mol	STD
1	4.8	4025	417	9740	394
2	5.0	4402	468	9488	350
3	5.2	4723	500	9680	405
4	5.5	4612	471	9717	406

N1 - the total number of conformation generated. N2 - the number of accepted conformations. E - the mean value of the energy of the accepted conformations, STD - Standard deviation of the energy of accepted conformations.

Before energy minimisation the model with an initial radius of 5.0 Å results in an ensemble of structures generated by the MCPBA method which have the lowest average energy (E = 9488 kcal/mol). The model with an initial radius of 4.8 Å results in an ensemble of structures which have the highest mean energy (E = 9740 kcal/mol). The second lowest energy ensemble (model 3) had an initial radii of 5.2 Å. The backbone structure of the coiled-coil is fixed during the simulated annealing procedure and therefore, the radius and pitch of the coiled-coil does not change. The mean structure of the ensembles is subjected to minimisation without fixing any atoms by the ***four*** step sequence outlined in the methods section. The models of the rod domain are each subjected to 3000 iterations of energy minimisation. The conformational energies of the mean minimised models are significantly different (step 2 methods section). The lowest energy structure (-22490 kcal/mol) for the total rod domain for these models is where the initial radius was set at 5.0 Å (model 2) and this model is close in energy to model 3 with an initial radii of 5.2Å. The difference between the lowest (-22490 kcal/mol) and the highest (-22232 kcal/mol) conformation energy is about 258 kcal/mol, an average difference of 0.05 kcal/mol for each atom in the model of rod domain.

The heptad repeats, radius and pitch and conformational energy of each minimised model are given in Table 5.7. The radii increase and the pitch is reduced by energy minimisation. The average pitch of the four models after minimisation is 188 Å and the heptad repeat, defined by the criteria given in section 5.3.1, in models 1 and 4 do not

change. In the model 2, the heptad repeat in segment 1A is lost and in model 3, the heptad repeat of segment 1B is lost. The heptad repeats in other segments do not change.

Table 5.7. Radius and pitch of different models after energy minimisation.

	Before EM	After EM		
Model	[a]Radius Å	[b]Radius Å	[c]Pitch Å	Energy kcal/mol
1	4.80	5.69	176	-22232
2	5.00	5.78	197	-22490
3	5.20	5.90	184	-22459
4	5.50	6.27	196	-22445

In the next step (step 3 methods section) the helical coiled-coil segments 1A, 1B, 2A and 2B of the four models were separated and separately minimised without constraints on any atoms for up to 1000 iterations. The link segments L1, L12 and L2 are not at this stage further minimised. The conformation energy, the radii, pitch and the heptad repeats in each of four helical segments for each of the four models were determined (Table 5.8).

Table 5.8. Energy minimisation of the four helical segments in models 1-4.

Model		1A	1B	2A	2B
1	[a]E (kcal/mol)	-1748	-4838	-959	-5266
	[b]r_o (Å)	5.18	5.80	5.23	5.57
	[c]P (Å)	174.7	203.2	152.1	190.7
	[d]Hr	[e]retained	retained	retained	retained
2	[a]E (kcal/mol)	-1498	-4945	-882	-5584
	[b]r_o (Å)	5.43	6.24	5.46	5.63
	[c]P (Å)	193.2	192.6	253.6	185.2
	[d]Hr	[f]changed	retained	retained	retained
3	[a]E (kcal/mol)	-1692	-4578	-877	-5512

	[b]r_o (Å)	5.39	5.76	5.88	5.67
	[c]P (Å)	169.5	203.8	146.8	184.7
	[d]Hr	retained	[g]lost	retained	retained
4	[a]E (kcal/mol)	-1685	-4886	-883	-5521
	[b]r_o (Å)	6.58	6.56	5.16	5.83
	[c]P (Å)	190.2	202.3	192.4	188.6
	[d]Hr	retained	retained	[f]changed	retained

a Energy of each helical segment, b Radius of the coiled-coil segment, c Pitch of the coiled-coil segment, d Heptad repeat, e Retained heptad repeat, f Different heptad repeat after energy minimisation. g No heptad repeat after energy minimization.

The lowest conformational energy of each of the four coiled-coil helical segments from each of the four models are: -1748 kcal/mol for 1A in the model 1, -4945kcal/mol for 1B in the model 2, -959 kcal/mol for 2A in the model 1 and -5584 kcal/mol for 2B in the model 2 respectively. The heptad repeats of the residues in segment 1A of model 2 and segment 2A of model 4 were changed by minimisation and the heptad repeats of the residues in segment 1B of model 3 were totally lost. The final coiled-coil model is assembled by combining the structural pieces of the rod domain selected from the helical segments which have the lowest conformational energy without a change in the heptad repeat from the initial structure (step 4; method section). The helical segments 1A and 2A in model 1 and the segments 1B and 2B in model 2 were selected for combination to give the final model of the rod domain. The link segments of L1, L12 and L2 of model 1 were selected to link the helical segments 1A, 1B, 2A and 2B respectively. The combined model was subjected to 300 energy minimisation iterations. The heptad repeat remained intact. The energy of the final structure of the coiled-coil rod domain was -22567 kcal/mol. The heptad repeats, radius and the pitch in each of four helical segments in the rod domain of this model have been examined.

5.3.3. Analysis of the model

The pictures of the full atom model of the rod domain, the C^{α} coordinates of the rod domain and the helical segments 1A, 1B, 2A and 2B are given in the Appendix 3.

5.3.3.1. Parameters in the coiled-coil rod domain

a) Parameters of single α-helix. The parameters for the single helical strand within the four helical segments are given in Table 5.9. The average rise per residue is 1.464 Å, rotation angle per residue is 102.999°, the number of residues per turn is 3.524 and the pitch is 5.171 Å. In a single stranded straight α-helix[44] these parameters are 1.50 Å for the rise per residue; 100° for the rotation angle of a residue; 3.6 for the number of residues per turn and 5.4 Å for the pitch.

Table 5.9. The average parameters of the α-helices in the final model of the rod domain.

Helical Segment	Rise per residue (Å)	Rotation angle (°)	Number of residues per turn	Pitch (Å)
1A	1.450	102.755	3.513	5.129
1B	1.487	103.595	3.506	5.231
2A	1.458	102.726	3.523	5.157
2B	1.462	102.920	3.521	5.167
Average	1.464	102.999	3.524	5.171

b) Parameters of the coiled-coil in the four segments. The coiled-coil results in distortion of the straight α-helix. The average value of the radius for the coiled-coil is 5.56 Å (Table 5.10) almost identical with the value of 5.5 Å suggested by Crick[14] and Fraser.[45] The radius varies for the different helical segments within the range of 5.24 to 5.92 Å. The average value of pitch for the four helical segments is 172 Å, lower than the value of 186 Å suggested by Crick.[14] In our model the pitch for the different segments varies between 124 and 196 Å. The crossing angles of the two coiled-coil chains in a final model are calculated according to the relation between the radius r_o, pitch P and the helical angle α as described by $\tan\alpha = 2\pi r_o/P$ such that the crossing angle is twice the

helical angle. The crossing angle varies for the different helical segments and has an average value of 23° (range 21° to 30°) (Table 5.10).

Table 5.10. The parameters of the four coiled-coil helical segments in the final model.

Segment	Radius (Å)	Pitch (Å)	Crossing angle (°)	Residues per major turn	Minor turns per major turn	Length (Å)
1A	5.24	177	21.07	129.6	36.9	49.8
1B	5.72	191	21.31	137.4	39.2	155.2
2A	5.36	124	30.38	98.3	27.9	25.6
2B	5.92	196	21.49	143.7	40.8	182.4
Average	5.56	172	23.56	127.3	36.2	

The number of residues per major helical turn is calculated from the rise per residue in one minor turn (Table 5.9) and the pitch (Table 5.10). The average number of residues per major helical turn in the model is 127 close to that (126) suggested by Crick.[14] The average value of the number of minor turns in a major coiled-coil turn is 36. The total length of the helical segments in the rod domain, not including linking regions, is 413 Å. This was measured using PCMODEL.[46] The length of the rod domain including helical segments and the linking segments is 472.6 Å close to the length (470 Å) previously suggested by Parry and Fraser.[5]

5.3.3.2. The distribution of the residues

(a) The distribution of the charged residues on the chains 8c-1 and 7c.

The distribution of residues in the rod domain of chain 7c[47] and 8c-1[48] of wool protein is given in Table 5.11.[49]

Table 5.11. Distribution of non-polar, polar, acidic and basic residues in the rod domain.

		N^e	Non-polar	Polar	Charged+ (Basic)	Charged- (Acidic)
Chain	Seg.		N^a	N^b	N^c	N^d
7c	1A	35	15	8	7	5
	1B	101	41	23	14	23
	2A	19	8	5	2	4
	2B	121	47	32	20	22
	linking	35	9	16	4	6
Overall		311	120	84	47	60
8c-1	1A	35	12	8	7	8
	1B	101	37	29	14	21
	2A	19	7	6	2	4
	2B	121	39	49	12	21
	linking	35	14	12	3	6
Overall		311	109	104	38	60

The number of residues in the helices: *a* non-polar, *b* polar, *c* positively charged residues, *d* negatively charged. *e* total number.

There are 120 and 109 non-polar residues (39% and 35% of all residues) and 84 and 104 polar residues (27% and 33% of all residues in rod domain) in chain 7c and 8c-1 respectively. There are 47 (15.1%) positively charged residues (basic residues) in chain 7c and 38 (12.2%) in chain 8c-1. The chain 7c is 'basic or neutral' relative to chain 8c-1. There are 60 (19%) negatively charged residues (acidic residues) in each of the chains 8c-1 and 7c.

The % distribution of the charged residues in each of the four helical segments in the two chains is given in Table 5.12.

Table 5.12. The % distribution of the charged residues in the four helical segments of the coiled-coil rod domain.

Segment	Non-polar %	Polar %	Charged$^+$ %	Charged$^-$ %	Exposed %
1A	38.6	22.8	20.0	18.6	98
1B	38.6	25.7	13.9	21.8	86
2A	39.5	28.9	10.5	21.1	100
2B	35.5	33.5	13.2	17.8	69

Segment 1A has the highest percentage of the positively charged residues in the four segments and is 'basic' relative to the segment 1B, 2A and 2B. 20% of the residues are positively charged in segment 1A and 18.6% negatively charged. Segment 1B has 13.9% of the residues positively charged and 21.8% negatively charged. Thus segment 1B is 'acidic' relative to the segment 1A. In segment 2A, the positively charged residues make up only 10.5% and negatively charged residues 21.1%. The segment is more 'acidic' than segment 1B. In segment 2B, the positively charged residues make up 13.2% and negatively charged residues 17.8%. The segment is more 'basic' than segment 1B and more 'acidic' than segment 1A.

(b) Core and exposed residues.

If a charged residue (positive or negative) is located at either the *a* or *d* positions of the heptad repeat, the residue is said to be in the 'core' of the inter-chain or 'buried' and if located at other than an *a* or *d* position is exposed and can interact with solvent. The percentage of charged residues exposed will influence the surface area and conformation of the coiled-coil. In the four coiled-coil segments of wool protein, segment 1A has only one charged residue (35R in chain 8c-1) which is in a *d* position and no charged residues in an *a* position. Thus 98% (69/70) of the charged residues are exposed. In segment 1B, 86% of the charged residues (124/144) are in other than *a* or *d* positions. In the segment 2A, all of the charged residues (30/30) are exposed. In segment 2B, 69% (104/150) of the charged residues are located at other than *a* or *d* positions.

5.3.3.3. The distribution of the backbone dihedral angles

The distribution of the backbone dihedral angles ϕ and ψ of the coiled-coil rod domain of the model of wool protein has been calculated using the program *tors*.[24] In the coiled-coil model, the dihedral angles of the backbone are distorted from the values in a straight helical chain and for the four helical coiled-coil segments are shown in Figure 5.4a. The mean value and standard deviation of the 552 dihedral angles, ϕ and ψ, in the four helical segments of the coiled-coil model of rod domain is -60° ± 27° and -40° ± 25° respectively. However, 5.6% (31/552) of the values of ϕ and 3.8% (23/552) of the values of ψ deviate more than 3σ from the mean values. The total number of the residues in the linking segments is 70 with 35 per chain. The dihedral angles in the linking segments scatter in the range of -180° to 180° (Figure 5.4b)

(a) (b)

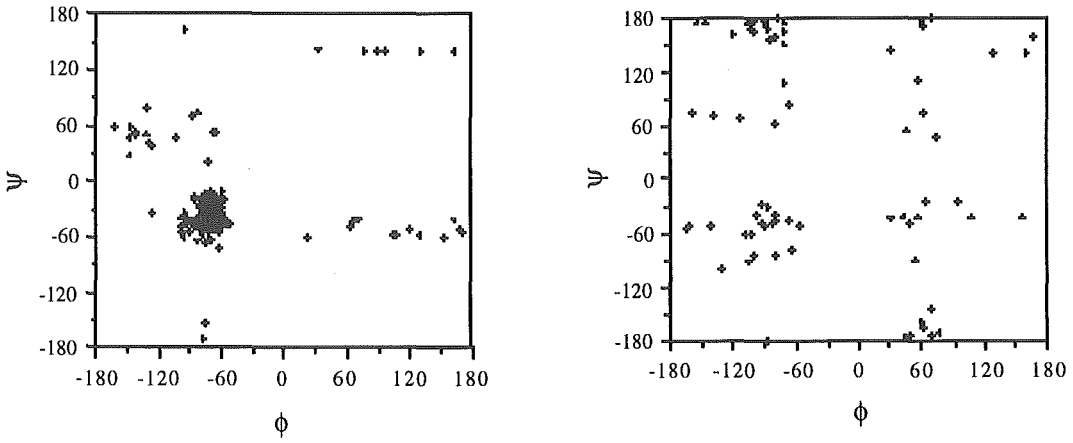


Figure 5.4. Distribution of the backbone dihedral angles in the (a) coiled coil and (b) linking segments of the rod domain.

5.3.4. The inter-chain interaction energy of the rod domain

The van der Waals and electrostatic interaction energy between the two coiled-coil chains of the model have been calculated using the program *mtranscoil*. The interaction energy terms take the form:[50]

$$E_{vdw} = F_{ij} / r_{ij}^{12} - C_{ij} / r_{ij}^{6}$$

$E_{el} = 332.0q_iq_{jj}/Dr_{ij}$

where F_{ij} and C_{ij} are the vdW energy parameters, r_{ij} is the distance between atoms *i* and *j*, 332.0 is a transform factor, q_i and q_j are partial charges on the atoms *i* and *j*, D is a constant, generally taken as 2.0. In the program *mtranscoil*, E_{vdW} and E_{el} non-hydrogen energy parameters can be selected from one of three force fields: ECEPP/2, OPLS and AMBER as developed by Scheraga,[51] Jorgensen[33] and Kollman[52] respectively. In this study we used the parameters of the OPLS force field. The cut-off distance between a pair of atoms for the van der Waals potential and electrostatic interactions is set at 6.0 Å and 8.0 Å respectively.

The E_{vdW} and E_{el} energy of interaction of the side-chains and backbone atoms within the above criteria in the rod domain including coiled-coil and linking segments in the final model is calculated for 16616 interactions.The major contribution to the inter-chain energy is from the electrostatic interactions. The inter-chain energy is calculated as -1554.1 kcal/mol in which the vdW energy is -51.1 kcal/mol and electrostatic energy is -1503.0 kcal/mol. The hydrogen bonds, disulfide bonds, non-polar interactions, polar and ionic interactions, heptad repeat interaction are also important and are discussed below.

5.3.4.1. Hydrogen bonds

a). Hydrogen bonds in the backbone of the coiled-coil rod domain.

The geometrical parameters of a hydrogen bond are defined in Figure 5.5:

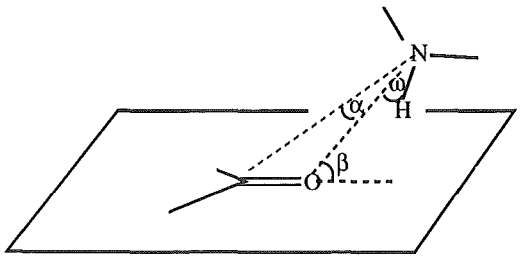


Figure 5.5. The parameters defining a hydrogen bond.

For hydrogen bonds in crystal structures of proteins the distance between an acceptor atom (X) and a donor atom (Y) is in the range 2.89 to 3.5 Å.[53,54,55,56] The angle α is in the range 0° to 50° and β, which is formed by an amide plane containing backbone atoms $C\alpha_i$, C'_i, O_i and line O_i-N_{i+4} is in the range 0° to 50°. For an α-helical structure the N...O distance is 2.95 Å, and α and β have values of 18° and 27° respectively. Calculation of the energy of the hydrogen bond is based on the expression developed by Scheraga et al.[57]

$$E_{H\text{-}B} = A_{Y\ldots H}/r_{Y\ldots H}^{12} - B_{Y\ldots H}/r_{Y\ldots H}^{10}$$

where $A_{Y\ldots H}$ and $B_{Y\ldots H}$ are parameters with a value of 12040 kcal A^{12}/mol and 4014 kcal A^{10}/mol respectively[51] and $r_{Y\ldots H}$ is the distance between the hydrogen atom and the acceptor atom. Because the models are not generated with hydrogen atoms the distance $r_{Y\ldots H}$ in the coiled-coil models can not be obtained directly. The angle of HYX (ω) has been reported to be in the range 0 to 20° in the hydrogen bonding models[57] and analysis of crystal structures[58] shows that the hydrogen atom of a hydrogen bond is close to the line between a donor and acceptor (Y ... X). To obtain a distance for $r_{Y\ldots H}$, the hydrogen atom is assumed to lie on this line. The value of the angle ω can not be obtained from the model which does not include hydrogen atoms but ω is not a variable in the hydrogen bond energy expression.

For the purposes of comparison, we separately calculated the parameters of the hydrogen bonds in a straight standard helix[59] before energy minimisation and in the coiled-coil helix before and after energy minimisation for a structure with 33 residues in each chain along with those for 2ZTA (Table 5.13).

Table 5.13. The hydrogen bond energy and geometric parameters of a standard α-helix, a coiled-coil helic and 2ZTA as in the crystal structure.

Geometry	Straight helix[a]	2ZTA	Coiled-coil[b]	Coiled-coil[c]
Y ... X (Å)	2.85	3.03	2.91	3.24
Angle α (°)	13.87	27.49	14.00	31.53
Angle β (°)	4.27	22.50	4.69	26.17
E_{H-B} (kcal/mol)	-1.043	-0.787	-0.998	-0.557

a Straight helix with the pitch 5.4 Å, rotation angle 100 per residue, 3.6 residues per turn. b Coiled-coil before energy minimisation. c Coiled-coil after energy minimisation.

The average Y ... X distance of a hydrogen bond in the coiled-coil before energy minimisation is 2.91 Å and is somewhat longer than in a 'standard' straight chain helix (2.85 Å). The average value of α (14°) in the coiled-coil is larger than in a straight helix (13°). The average energy of a hydrogen bond in the straight helix is -1.043 kcal/mol. For the coiled-coil structure, the average energy is -0.998 kcal/mol before energy minimisation. Energy minimisation of the coiled-coil results in distortion of the hydrogen bonds. The average Y ... X distance increases to 3.24 Å and the angles α and β increase to 31° and 26° respectively from average values of 2.91 Å, 14° and 4° respectively before energy minimisation. The average hydrogen bond energy decreases from -0.998 kcal/mol before energy minimisation to -0.557 kcal/mol after energy minimisation. For the four coiled-coil helical segments of wool protein the angles α and β, the hydrogen bond length (Y ... X) and the energy of the hydrogen bonds do not differ significantly.

A hydrogen bond in the coiled-coil backbone is formed between O_i and N_{i+4} if the distance Y...X is less than 3.5 Å. If O_i is in the *a* position of a heptad, the atom N_{i+4} will be in the *e* position of the same heptad. Thus the heptad positions which can form a hydrogen bond are *a-e, b-f, c-g, d-a, e-b, f-c* and *g-d*. The hydrogen bonds parameters between these heptad positions vary markedly and differ for 7c and 8c-1 (Table 5.14).

Table 5.14. The average geometry of the hydrogen bonds in the heptad repeats of the final model of the helical segments of the rod domain.

Geometry	Chain	*a-e*	*b-f*	*c-g*	*d-a*	*e-b*	*f-c*	*g-d*	Ave[a]
Y ... X	7c	3.22	3.28	3.26	3.21	3.24	3.27	3.18	3.24
	8c-1	3.22	3.26	3.22	3.22	3.24	3.22	3.18	3.22
Angles α	7c	30°	31°	31°	31°	32°	33°	30°	31°
	8c-1	31°	33°	31°	33°	34°	31°	32°	32°
Angle β	7c	24°	27°	24°	25°	27°	27°	25°	25°
	8c-1	27°	28°	25°	25°	29°	25°	26°	26°
E(kcal/mol)	7c	-0.560	-0.473	-0.499	-0.578	-0.538	-0.474	-0.630	-0.536
	8c-1	-0.567	-0.512	-0.565	-0.572	-0.530	-0.565	-0.635	-0.563

The shortest Y ... X distance (3.18 Å) is between the *g-a* positions. The distances of Y ... X in the positions *a-e* and *d-a* are 3.21 and 3.22 Å respectively in chain 7c. The average value of the distance Y ... X is 3.22 Å for both *a-e* and *d-a* hydrogen bonds in chain 8c-1. The lowest average energy of the hydrogen bond is -0.635 kcal/mol for *gd* hydrogen bonds in chain 8c-1. In chain 7c, the lowest hydrogen bond energy is -0.630 kcal/mol between *g -d*. The varying values of α and β reflect the distortion of the coiled-coil. The hydrogen bonds in the coiled-coil are most distorted for the interactions of *b-f* and *e-b* positions in both chains.

Hydrogen bonds also form between the backbone and a side-chain and between backbone and solvent and side-chain and solvent. We have not included solvent in the calculation of the inter-chain interactions in our model. For the potential hydrogen bonds of the backbone atoms between residues four apart the distance between the nitrogen of the NH donor and the oxygen of the acceptor is in the range 2.7 to 7.0 Å. If the distance between a donor and acceptor is larger than 3.5 Å, the hydrogen bond is considered to be deformed[55] and is not computed in the energy calculation. For chain 7c, the longest distance of Y ... X (6.76 Å) is between 143Val and 147Asn. In the chain 8c-1, the longest distance is 7.66 Å between 47Tyr and 51Phe. Thirty-seven (14%) hydrogen bonds are longer than 3.5 Å in chain 7c and 61 (23%) in chain 8c-1.

The hydrogen bonds of chain 8c-1 are more distorted than for chain 7c. The average values of the angles α and β in chain 8c-1 are larger than in chain 7c. In chain 8c-1 the average Y... X distance is 3.223 Å and the average energy of the hydrogen bond is -0.564 kcal/mol, somewhat less than the corresponding values (3.237 Å and -0.536 kcal/mol) in chain 7c. The total hydrogen bonding energy of the backbone in the rod domain is -231.7 kcal/mol (-119.5 kcal/mol in chain 7c and -112.2 kcal/mol in chain 8c-1).

Table 5.15. The potential hydrogen bonds found in the backbone atoms of the coiled-coil rod domain.

Chain	Number of H-B's	Deformed (%)	Y ... X Distance	α	β	E (kcal/mol)
7c	260	14%	3.237	31°	25.6°	-0.536
8c-1	260	23%	3.223	32°	26.4°	-0.564

The hydrogen bonds in the linking segments L1, L12 and L2 are included.

b). Hydrogen bonds between the protein side-chains.

Hydrogen bonds between protein side-chains are important in defining packing and folding.[60,61] The distance between the side-chains of the donor (Y) and acceptor (X) (Y ... X) range from 2.5 to 3.5 Å in fifty protein high resolution crystal structures[56] and similarly for the side-chains in the inter-chain region of the coiled-coil. In the final model of the rod domain of wool protein there are 35 potential interchain hydrogen bonds between the side-chains of 8c-1 and 7c. The hydrogen bonds with a Y ... X distance <3.0 Å are reported in Table 5.16.

Table 5.16. The strongest inter-chain hydrogen bonds in the coiled-coil rod domain.

N_t	N_1	R_1	Hpt_1	Atm_1	Atm_2	Hpt_2	R_2	N_2	Distance (Å)	Energy kcal/mol
1	36	Q	/	OE1	NE	*d*	R	35	2.93052	-1.08958
2	36	Q	/	OE1	NH1	*d*	R	35	2.97787	-1.02193
3	36	Q	/	OE1	NH2	*d*	R	35	2.87187	-1.09405
4	294	D	*a*	OD1	NH1	*g*	R	292	2.95478	-1.05951
5	294	D	*a*	OD1	NH2	*g*	R	292	2.94119	-1.07774

Note: Nt is the number of inter-chain hydrogen bonds. N_1 is the sequence number of residue in chain 1 (7c). R_1 is the residue in chain 7c. Hpt_1 is the heptad repeat position of the residues of chain 7c. Atm_1 - the atom name in the residue of chain 7c. Atm_2 - the atom name in the residues of the chain 8c-1. Hpt_2 is the heptad repeat position of the residues in the chain 8c-1. R_2 is the residues in chain 8c-1. N_2 is the sequence number of the residues in chain 8c-1. The nomenclature and the symbol of the atoms in an amino acid residue is based on system used in the PDB file. The sequence number in the coiled-coil rod domain is from 1 to 311 for each chain.

The strongest hydrogen bonds between side-chains are formed between glutamine (Q) and arginine (R), and between aspartic acid (D) and arginine (R). There are three hydrogen bonds between the side-chain of residue 36Q of chain 7c and residue 35R of chain 8c-1. The residue 36Q is in the L1 region and residue 35R is the last residue of the helical segment 1A and is in the *d* position of the heptad repeat. The oxygen atom OE1 of the amino group of the glutamine side-chain forms three hydrogen bonds with nitrogen atoms NE, NH1 and NH2 of the arginine side-chain. Two hydrogen bonds occur between the side-chain of residue 294D of chain 7c and residue 292R of chain 8c-1. Oxygen atoms of the carbonyl group of the aspartic acid side-chain form two hydrogen bonds with the nitrogen atoms, NH1 and NH2, in the arginine side-chain. These hydrogen bonding interactions occur between the positions *a* and *d* of the heptad repeat. The stereochemistry of the hydrogen bonds formed between 36Q and 35R, and between 294D and 292R are shown in Figure 5.6.

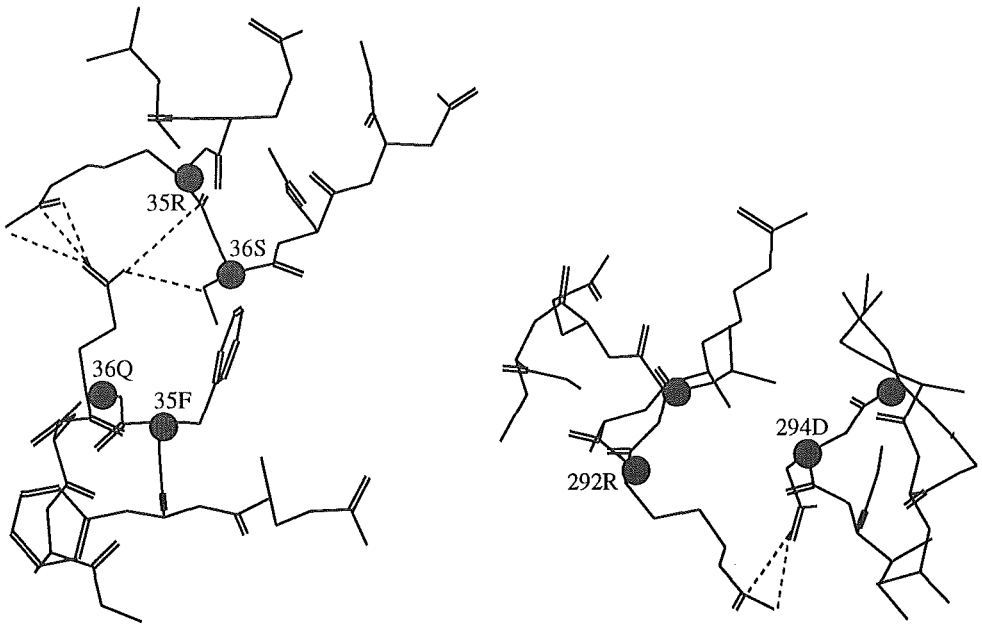


Figure 5.6. Stereochemistry of hydrogen bonding. (a) between residue 36Q in chain 7c and residue 35R in chain 8c-1. The residue 36Q in chain 7c is in the linking segment L1 and not in a heptad repeat. 35R is in a *d* position of a heptad repeat. (b) between residue 294D in chain 7c and 292R in chain 8c-1. The residue 294D is in the *a* position and residue 292R in the *d* position of the heptad repeat.

Hydrogen bonding energy between side-chains is small in comparison with the total interaction energy between side-chains. The total energy of the hydrogen bonds between the side-chains of the coiled-coil rod domain is calculated as -20.0 kcal/mol, while the total interaction energy between the side-chains in the coiled-coil segments is -1554 kcal/mol. There are thirty-one potential hydrogen bonds in the four helical segments of the coiled-coil rod domain of wool protein. A further four hydrogen bonds are located in the linking regions between the helical segments. The hydrogen bond energy of the side-chains between two helical chains is -15.97 kcal/mol in the four helical segments and -4 kcal/mol in the linking segments. The hydrogen bond energies between the side-chains in the four helical segments of the rod domain of wool are reported in Table 5.17.

Table 5.17. Hydrogen bonds between side-chains in the coiled-coil rod domain.

Segments	Number of H-B	Energy (kcal/mol)
1A	3	-1.353
1B	4	-1.531
2A	1	-0.662
2B	23	-12.431

5.3.4.2 Disulfide bonds in the coiled-coil rod domain.

Cysteine residues are unique in that they are capable of forming covalent disulfide linkages within or between polypeptide chains. Fraser et al[62] have investigated the distribution of disulfide bonds in α-keratin and concluded that there were no disulfide bonds between the chains of the coiled-coil. It is known that there are disulfide bonds between pairs of coiled-coil structures giving rise to the tetrameric structures. The nature of the disulfide linkages found in globular proteins has been discussed by Scheraga et al.[50] and experimental values[63] of bond lengths, bond angles, and the distance between two Cβ atoms of two cysteines involved in a disulfide bond are given in Figure 5.7.

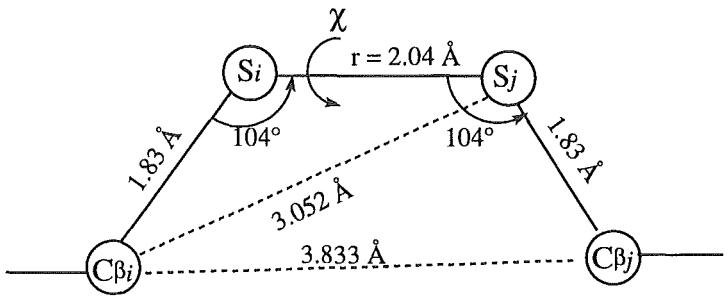


Figure 5.7. Experimental parameters of a disulfide bond.

The sequences[26,28] including N- and C-termini of components 7c and 8c-1 of the rod domain contain 56 cysteine residues. There are 8 cysteine residues in both the N-terminal domain and the rod domain and 9 in C-terminal domain of chain 8c-1. There are 10 cysteine residues in both the N-terminal and the rod domain and 11 in the C-terminal of chain 7c (Table 5.18).

Table 5.18. Number and locations of cysteine residues in the coiled-coil rod domain of 7c and 8c-1.

Protein	Total	Residue Number in rod domain (from 1 to 311 in each chain)						
Chain	Number	1A	L1	1B	L12	2A	L2	2B
7c	10	6(*c*)	40, 41	62(*c*), 112(*c*)	-	167(*f*)	-	195(*f*), 235(*g*), 260(*e*), 278(*b*)
8c-1	8	-	44	63(*g*), 116(*d*), 133(*g*)	-	-	-	212(d), 253(*c*), 295(*c*), 309(*c*)

The distance between the C^{β} atoms of two proximate cysteines is used as a criterion to define a disulfide bond such that if the distance is less than 4.0 Å a disulfide bond can exist. In the coiled-coil rod domain of wool protein, the possibility for intra-chain disulfide bond formation is limited since the formation of such bonds requires the cysteine residues to be in *a-a* or *d-d* positions of the heptad repeats.[64] In the final model no two cysteine residues in the rod domain of wool meet this requirement, consistent with Fraser's[64] prediction. The shortest distance between two C^{β}'s of cysteine residues in the rod domain of the model is 4.44 Å and this occurs in the segment 1B between the cysteine residue 62(*c*) in chain 7c and the cysteine residue 63(*g*) in chain 8c-1. These residues therefore do not form a disulfide linkage. The distances between other pairs of cysteine residues in chain 7c and 8c-1 are all greater than 5.0 Å, too far removed to form a disulfide bond.

5.3.4.3. Non-polar residue interactions of the rod domain.

It is generally accepted that the principal driving force in protein folding is the hydrophobic vdW interaction between non-polar side-chains.[65,66,67] Recently, Murphy et al.[68] have suggested that this effect is in fact destabilizing. We have calculated the vdW inter-chain interactions of the non-polar residues Gly, Ala, Val, Leu, Ile, Pro, Phe and Met. in our model to see if these interactions are stabilising or destabilising. Aromatic-

aromatic interactions are not included and are considered separately as are interactions of non-polar and polar residues, and non-polar and ionic residues.

i. Interactions between non-polar residues and non-polar residues.

In the final model a total of 2742 pairs of non-polar-non polar interactions aromatic were found in the interface of the coiled-coil chains and linking segments in the rod domain. The interaction energy is 5.12 kcal/mol, made up from a vdW energy of -5.82kcal/mol and an electrostatic energy of 10.94 kcal/mol. Only one interaction, namely between 262L of chain 7c and 261L of chain 8c-1,was within 3.5 Å. The non-polar - non-polar interactions excluding aromatic - aromatic interactions are therefore somewhat destabilizing.

ii. Interaction between non-polar residues and polar residues.

A total of 2398 interactions of non-polar and polar residues occur in the coiled-coil and linking segments of the rod domain. The interchain interactions are most common between *a* and *d* and *g* and *d* positions in the 7c and 8c-1 chains. Interactions with a distance less than 3.5 Å are shown in Table 5.19. The energy of these all interactions is -161.2 kcal/mol. The vdW interaction is -5.7 kcal/mol and the electrostatic interactions -155.5 kcal/mol. The interaction energy between non-polar and polar residues is greater than between non-polar and non-polar residues.

Table 5.19. The interaction between non-polar and polar residues in the rod domain.

N_t	N_1	R_1	Hpt_1	Atm_1	Atm_2	Hpt_2	R_2	N_2	Distance
1	25	N	*a*	ND2	CD1	*d*	L	28	3.47996
2	136	L	*g*	C	OD1	*d*	N	137	3.44610
3	136	L	*g*	O	OD1	*d*	N	137	3.48644
4	136	L	*g*	O	ND2	*d*	N	137	3.47509
5	220	N	*d*	OD1	CD1	*d*	L	219	3.40438
6	230	V	*g*	O	NE2	*d*	Q	233	3.46738
7	238	S	*a*	O	CD2	*d*	L	240	3.35432
8	276	M	*d*	O	NE2	*a*	Q	279	3.28469
9	276	M	*d*	CG	NE2	*a*	Q	279	3.49160
10	280	L	*a*	CD2	OE1	*a*	Q	279	3.48652

iii. Interaction between non-polar and ionic residues.

A total of 3502 interactions between non-polar and ionic residues in the rod domain have been evaluated. The total interaction energy is -342.7 kcal/mol. The vdW interaction energy is -5.8 kcal/mol and the electrostatic interaction -337.1 kcal/mol. There are four interactions with a contact distance of less than 3.5 Å. Three are between the non-polar residue 276M in chain 7c and the charged residue 275D in chain 8c-1. Both residues 276M and 275D are in *d* positions. The fourth interaction occurs between the charged residue 266E in chain 7c and non-polar residue 265V in chain 8c-1. The distance between the oxygen atom of the carbonyl in glutamic acid (E) and the nitrogen of the backbone of valine (V) is 3.39Å.

Table 5.20. The major interactions of non-polar and ionic residues.

N_t	N_1	R_1	Hpt_1	Atm_1	Atm_2	Hpt_2	R_2	N_2	Distance Å)
1	266	E	*a*	OE1	N	*a*	V	265	3.39872
2	276	M	*d*	CG	OD1	*d*	D	275	3.42252
3	276	M	*d*	CG	OD2	*d*	D	275	3.33175
4	276	M	*d*	CE	O	*d*	D	275	3.45747

5.3.4.4. Polar residue interactions in the rod domain.

The interactions of polar with polar and polar with ionic residues are considered separately.

i. Interactions between polar with polar residues

The polar residues, Asn, Gln, Cys, Thr, Ser, Tyr, and Trp of the rod domain of wool give rise to 575 interactions. The inter-chain interaction energy of the polar residues Tyr and Trp are calculated separately. The total energy of the polar-polar interactions in the rod domain including linking segments is -73.8 kcal/mol made up of a vdW energy of -3.2 kcal/mol and an electrostatic energy of -70.6 kcal/mol. The interaction of the polar residues with separations of less than 3.5 Å are listed in Table 5.21. 25N in chain 7c can interact with 25N in segment 1A of chain 8c-1. These residues are in the *a* position of the heptad repeat. There is a polar interaction between residue 152S and residue 148Q in the link segment L12. A similar interaction in the 2B segment occurs between residues 235C and 233Q. Residue C is in the *e* position and Q in the *d* position of the heptad repeat.

Table 5.21. Major interactions between polar - polar residues.

N_t	N_1	R_1	Hpt_1	Atm_1	Atm_2	Hpt_2	R_2	N_2	Distance (Å)
1	25	N	*a*	OD1	CA	*a*	N	25	3.28085
2	25	N	*a*	OD1	CB	*a*	N	25	3.23534
3	25	N	*a*	OD1	ND2	*a*	N	25	3.12417
4	152	S	/	OG	NE2	/	Q	148	3.11560
5	235	C	*e*	N	OE1	*d*	Q	233	3.09654

ii. Interaction between polar and ionic residues .

There are 1789 interactions between polar and ionic residues in the rod domain resulting in an interaction energy of -358.8 kcal/mol of which the vdW contribution is -11.0 kcal/mol and the electrostatic interaction -347.8 kcal/mol. The residues within 3.5 Å are listed in Table 5.22.

Table 5.22. Interactions between polar and ionic residues in the rod domain.

N_t	N_1	R_1	Hpt_1	Atm_1	Atm_2	Hpt_2	R_2	N_2	Distance (Å)
1	25	N	*a*	ND2	OE2	*g*	E	24	3.27672
2	36	E	/	OE2	N	/	S	36	3.18808
3	36	E	/	OE2	CB	/	S	36	3.34247
4	140	E	*d*	OE1	CA	*d*	N	137	3.20767
5	140	E	*d*	OE1	CG	*d*	N	137	3.22405
6	173	K	*a*	CD	OG1	*d*	T	172	3.48850
7	231	E	*a*	CA	NE2	*d*	Q	233	3.40742
8	231	E	*a*	OE2	CD	*d*	Q	233	3.37790
9	234	K	*d*	CB	OE1	*d*	Q	233	3.40894
10	252	E	*a*	OE2	CB	*d*	Q	254	3.38006
11	259	R	*a*	CD	O	*g*	Q	257	2.89585
12	259	R	*a*	NE	O	*g*	Q	257	3.43298
13	259	R	*a*	NE	OE1	*g*	Q	257	3.38831
14	259	R	*a*	CZ	OE1	*g*	Q	257	2.89565
15	259	R	*a*	NH1	OE1	*g*	Q	257	3.17529
16	266	E	*a*	CA	OE1	*d*	Q	268	3.40585
17	266	E	*a*	CB	OE1	*d*	Q	268	3.47675
18	266	E	*a*	CG	OE1	*d*	Q	268	3.40903
19	266	E	*a*	OE1	C	*g*	N	264	3.34624
20	266	E	*a*	OE1	CB	*g*	N	264	3.33242
21	266	E	*a*	OE1	CG	*g*	N	264	3.33537

This class of interactions also includes hydrogen bonds between the polar side-chain and the backbone. There are 14 hydrogen bonds between polar residue side-chains and ionic residues. Hydrogen bonds can also form between the nitrogen atom of the side-chain of residue 259R and the oxygen atom of side-chain 257Q. The stereochemistry of the interaction between residue 259R in chain 7c and 257Q in 8c-1 and between 25N in 7c and 24E in 8c-1 is shown in Figure 5.8.

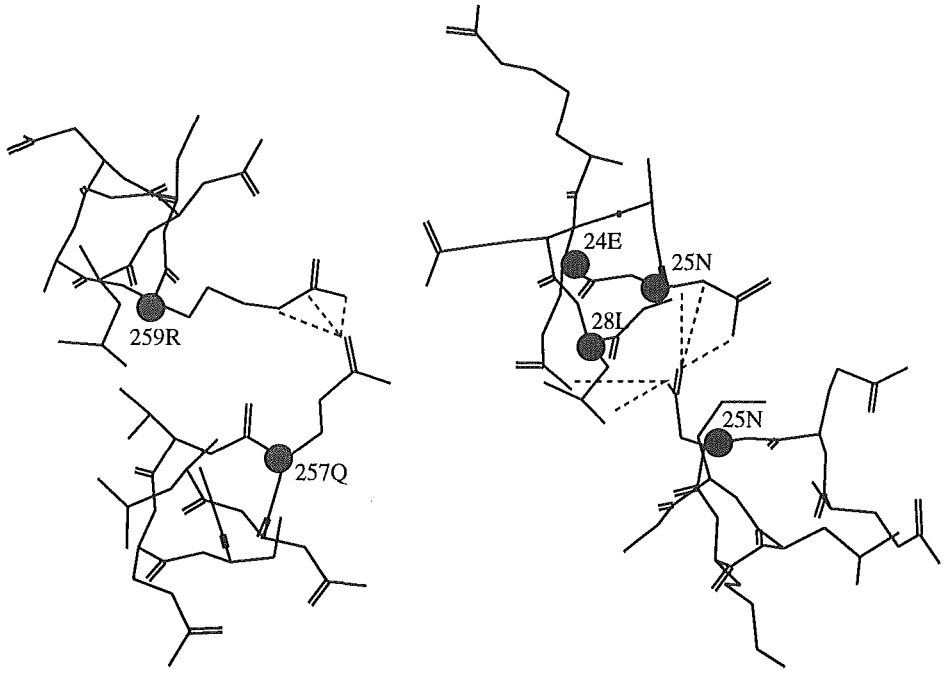


Figure 8. Stereochemistry of the strongest interactions between polar-ionic residues. (a) between residues 259R in chain 7c and 257Q in chain 8c-1. 259R is in the *a* position and the residue 257Q is in the *g* position. Three nitrogen atoms of arginine residue in the *a* position can interact with the oxygen of glutamine residue in the *g* position. (b) between residues 25N in 7c and 24 E in 8c-1. The residue 25N is in the *a* position and 24E is in the *g* position. The interaction distance between the nitrogen atom of the side-chain of 25N and the oxygen atom of side-chain of 24E is 3.27 Å. The residue 25N in a *a* position of 7c is also in close contact with 28L in a *d* position in 8c-1 and 25N of the *a* position in 8c-1 (see Tables 5.19 and 5.21).

5.3.4.5. Ionic interactions in the rod domain.

Charged side-chains have an important role in stabilising α-helices[69,70] and have been postulated to be important in determining alignment in fibrous proteins.[71] The charged residues occupy *b, c, g, f* positions in the heptad repeat. In our model, there are 1705 interchain interactions within the cut off criteria between the ionic-ionic residues in the coiled-coil rod domain and these interactions contribute -302.1 kcal/mol of interaction energy. The ionic-ionic interactions where the distance is less than 3.5 Å are reported in Table 5.23.

Table 5.23. Ionic interaction between two coiled-coil chains.

N_t	N_1	R_1	Hpt_1	Atm_1	Atm_2	Hpt_2	R_2	N_2	Distance (Å)
1	36	E	/	OE1	CD	*d*	R	35	3.39688
2	36	E	/	OE1	CZ	*d*	R	35	2.57123
3	36	E	/	OE2	C	*d*	R	35	3.43027
4	36	E	/	OE1	NH1	*d*	R	35	2.97787
5	36	E	/	OE1	NH2	*d*	R	35	2.87187
6	140	E	*d*	CA	OE2	*a*	E	141	3.25360
7	173	K	*a*	NZ	CA	*e*	R	173	3.38984
8	241	E	*d*	OE2	CD	*a*	R	237	3.31540
9	252	E	*a*	OE1	C	*g*	R	250	3.45397
10	294	D	*a*	OD1	NE	*g*	R	292	3.04618
11	294	D	*a*	OD1	NH1	*g*	R	292	2.95478
12	294	D	*a*	OD1	NH2	*g*	R	292	2.94119
13	294	D	*a*	CG	NE	*g*	R	292	3.49290
14	294	D	*a*	CG	CZ	*g*	R	292	3.46376
15	294	D	*a*	OD1	CZ	*g*	R	292	2.63495
16	308	E	*a*	OE1	CA	*a*	E	307	3.29197

There are nine hydrogen bonds between the ionic side-chains in the rod domain where the distance between a N atom and a O atom is less than 3.5 Å. The interactions of ionic residues between the two charged atoms can form a ionic salt. This type of the interaction is not as strong as the interaction between a polar residue and an ionic residue.

5.3.4.6. Interaction involving aromatic residues

Aromatic π–π interactions are important in molecular recognition[72] and play an important role in controlling the conformation and substrate binding properties of nucleic acids and proteins[73,74] and are separately considered as a special group in our model.[75] along with the interaction energy between aromatic-aromatic and aromatic-non-aromatic residues. In the rod domain, there are 4128 interchain interactions between the aromatic and aromatic residues and non-aromatic residues in the coiled-coil rod domain. The total energy of these interactions is -255.3 kcal/mol of which the vdW energy is -8.9 kcal/mol

and the electrostatic energy -246.3 kcal/mol. The aromatic-aromatic and aromatic-non aromatic interactions where the contact distance between interacting atoms is less than 3.5 Å are listed in Table 5.24. The sum of non polar - non polar interactions when aromatic-aromatic interactions are included is stabilising.

Table 5.24. Inter-chain interactions involving aromatic residues

Nt	N1	R1	Hpt1	Atm1	Atm2	Hpt2	R2	N2	Distance (Å)
1	15	I	*e*	CA	CD2	*d*	Y	14	3.34871
2	15	I	*e*	CB	CD2	*d*	Y	14	3.39504
3	52	Y	*g*	CD2	O	*d*	T	53	3.25325
4	84	Y	*d*	CD2	O	*a*	F	85	3.18365
5	88	Y	*a*	CA	OE1	*a*	E	92	3.35934
6	88	Y	*a*	CD2	OG1	*g*	T	91	3.21271
7	137	Y	*a*	CA	OD1	*d*	N	137	3.43131
8	137	Y	*a*	CG	OD1	*d*	N	137	3.28015
9	137	Y	*a*	CD2	OD1	*d*	N	137	3.31941
10	137	Y	*a*	CD2	O	*g*	C	133	3.27815
11	137	Y	*a*	CD2	SG	*g*	C	133	3.42944
12	176	Y	*d*	O	CE1	*a*	Y	176	3.30057
13	248	E	*d*	O	CE1	*a*	Y	251	3.38592
14	252	E	*a*	OE1	CA	*a*	Y	251	3.48407
15	252	E	*a*	OE2	CA	*a*	Y	251	3.47875
16	283	Y	*d*	O	CD1	*a*	L	286	3.44218
17	283	Y	*d*	CB	OE2	*d*	E	282	3.49433
18	283	Y	*d*	CZ	OE1	*a*	Q	279	3.44197

The number of aromatic residues in the helical segment of the rod domain is 33 and 16 (6 Phe and 10 Tyr) are located at *a* or *d* positions. Of the remaining 17 aromatic residues, four histidine residues in the rod domain are located at position 137H(*e*) and 234H(*e*) in chain 8c1 and the 76H(*c*) and 205H(*c*) in chain 7c. A few aromatic residues are located in *b* and *g* positions. No aromatic residues are in the *f* position of the heptad repeat (Table 5.5). Most of the interactions involving aromatic residues occur between residues in *a, d, e* and *g* positions (Table 5.24). In the model, the seven tryptophan (Trp,

W) residues in the linkage segments were replaced by alanine residues to bring the atomic number of the rod domain to less than 5000, a requirement of Macromodel. This will limit the accuracy of the model of the linking region but is not expected to have a significant influence in the helical segments.

5.3.5. The interaction energy of the different type of the interaction

The interaction energy between the two chains in the coiled-coil rod domain is a stabilising interaction and different components to this energy are given in Table 5.25.

Table 5.25. The contribution of the various types of interactions of the side chains to the stabilisation energy of the rod domain.

Interaction type	Interactions	Contribution (%)	Energy (kcal/mol)	Contribution (%)	[a]Ep
Hydrogen bonding	35	0.2	-20	1.3	-0.57
Aromatic residues	4128	24.5	-255	16.8	-0.06
Nonpolar-nonpolar	2742	16.2	5	-0.3	0.002
Nonpolar-polar	2398	14.2	-161	10.6	-0.07
Nonpolar-ionic	3502	20.7	-343	22.6	-0.10
Polar-polar	575	3.4	-74	4.9	-0.13
Polar-ionic	1789	10.6	-359	23.7	-0.20
Ionic-ionic	1705	10.1	-302	19.1	-0.17

[a] average energy (kcal/mol) per interaction.

The most important interaction (-359 kcal/mol) occurs between polar residues and ionic residues and contributes 23.7% to the total interaction energy. The next most important interaction is between non-polar residues and ionic residues (22.6%). The interaction between ionic residues and ionic residues contributes 19.1% to the total interaction energy. Aromatic-aromatic interactions contribute 16.8% and polar-polar interactions 4.9%. Hydrogen bonding interactions between the side-chains contribute <2% to the interaction energy. The interaction energy between non-polar and non-polar residues has a negative contribution (-0.3%) to the total interaction energy. Aromatic residues are associated with 24.5% of the interactions. Non-polar-ionic interactions are

20.7% of the total numer of interactions. The interactions of ionic residues (ionic-nonpolar, ionic-polar and ionic-ionic residues) contribute 40% to the total number of interactions. The interaction energy involved with ionic residues is ca. 70% of the total for the coiled-coil rod domain. The interaction energy involved in polar residues (polar-polar, polar-nonpolar and polar-ionic residues) is 39%. The number of such interactions is about 18% of all interactions. This type of interaction is not as strong as interactions involving ionic residues. The interaction energy involving non-polar residues (nonpolar-nonpolar, nonpolar-polar, and non-polar-ionic residues) contribute 32% of the interaction energy. Interactions involving non-polar residues between two chains made up 51% of the total number of interactions.

While it is reasonable that most of the non-polar residues are located in the core or inter-face between the coiled-coils the interaction energy associated with these interactions is not large compared with other types of interactions. Excluding aromatic-aromatic interactions the non-polar non-polar interactions are somewhat destabilising. Some 16% (-255 kcal/mol) of the interaction energy is from interactions involving the aromatic residues. There are 4128 such interactions or 25% of the total number of interactions. The greatest number of interactions involving aromatic residues involve interactions with the non-aromatic residues. There are 839 aromatic-aromatic interactions corresponding to an interaction energy of -47.4 kcal/mol. The total interaction involving aromatic residues is less than the interaction energy involved with ionic residues but greater than the interaction energy associated with non-polar residues.

The interaction energy of the different types of interactions are shown as energy per interaction in Table 5.25. The strongest interaction is hydrogen bonding, (-0.57 kcal/mol per interaction). The interaction energy between polar residues and ionic residues is -0.20 kcal/mol per interaction. The ionic-ionic interactions have an energy of -0.17 kcal/mol and the average polar-polar interaction is -0.13 kcal/mol. The average aromatic-aromatic interaction energy is -0.06 kcal/mol. The interaction energy associated with ionic residues or with hydrogen bonding is greater than other interactions. From the calculation of inter-chain interactions, we conclude that the major interaction in the coiled-coil rod domain in wool protein is from charged residues. The aromatic-aromatic

interactions and polar-polar interactions also have an important role in establishing the stability of the coiled-coil structure. Hydrogen bonding interactions are strong, but the number of these interactions is low. The interactions between non-polar and non-polar residues have the greatest influence in destabilizing the coiled-coil conformation.

5.3.6. The interaction energy of the various heptad position

Most non-polar residues are located in the *a* and *d* position in the core or inter-face of the two coiled-coil chains resulting in the so called 'hole' and 'knob' interaction.[14] The residue at the *a* position of the ith heptad repeat of one chain can interact with the residues in the (i-1)*d*, (i-1)*g*, (i)*a*, and (i)*d* positions of other chain. The residue in the (i)*d* position of one chain can interact with the residues in the (i)*a*, (i)*d*, (i)*e* and (i+1)*a* positions of other chain[7] (Figure 5.9).

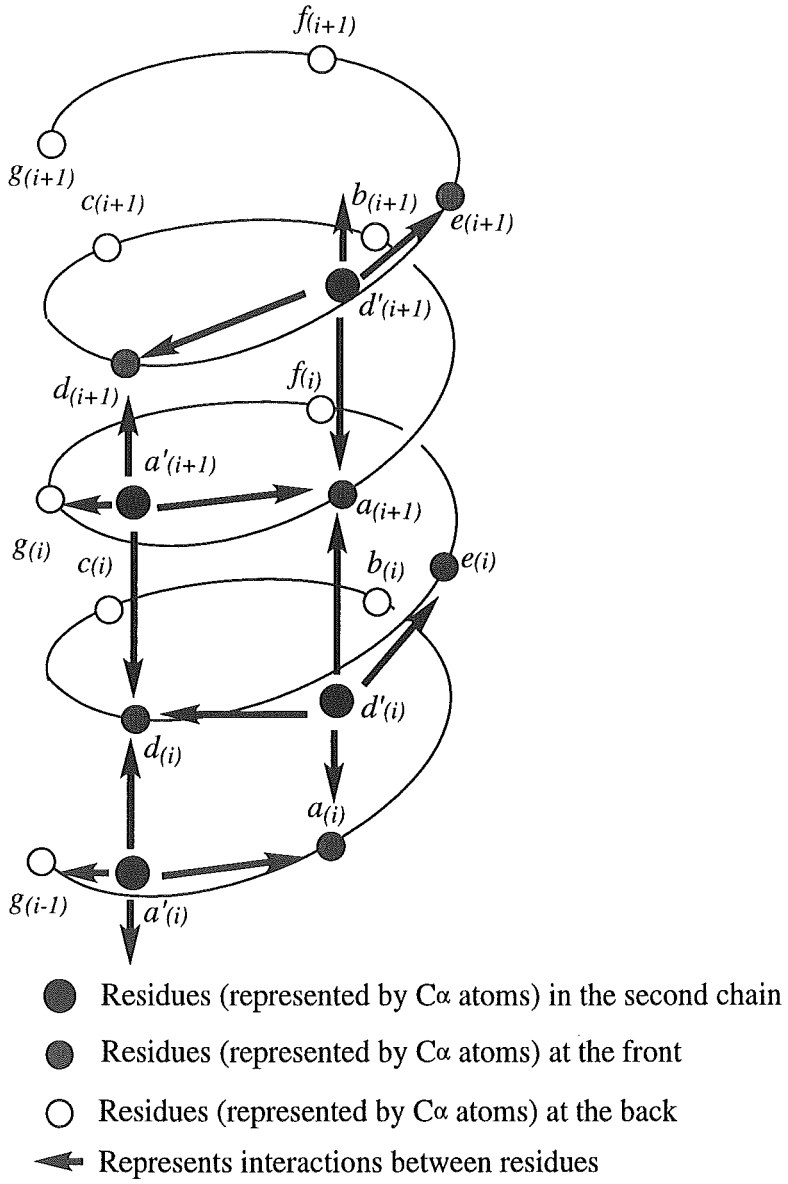


Figure 5.9. The interactions between the heptad positions of the two helices in a coiled-coil structure.

The energy of interaction between the various positions of the heptad for the rod domain of wool including the linking segment are given in Table 5.26. Heptad positions are only defined in the helical segments. The linking segments (L1, L12 and L2) have no heptad definition although they contribute to the energy of interaction between the two chains.

Table 5.26. Interaction energy in each of the coiled coil heptad repeat positions.

				Chain 7c				
8c-1	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*linkage*
a	1491/-131	97/-38	148/-84	1805/-187	298/-29	91/-27	1090/-113	1/5
b	89/-38	0/0	0/0	280/-48	6/10	0/0	49/7	0/0
c	249/-56	0/0	0/0	154/-29	0/0	0/0	0/0	0/0
d	2398/-161	195/-44	182/13	1952/-157	948/-22	108/-23	555/-61	57/2
e	349/-25	0/0	5/-9	738/56	23/41	2/9	246/19	0/0
f	118/-45	0/0	0/0	87/-7	0/0	0/0	3/9	0/0
g	1062/-116	19/19	7/-1	632/-66	255/-12	6/8	62/-3	0/0
linkage	40/-5	13/-23	18/-7	134/-31	136/-2	33/-4	9/-1	375/-37

Energy in kcal/mol. linkage - the interaction between linking segments. The heptad repeats in the linking segments have no definition.

The major interactions contributing to the interaction energy of the two coiled-coil chains is from *a-a* (-131 kcal/mol), *a-d* (-187kcal/mol), *d-a* (-161kcal/mol) and *d-d* (-157kca/mol) positions of 7c and 8c-1. The total interaction energy associated with *a* and *d* interactions is -636kcal/mol or some 41% of the total interaction energy between the chains. The interactions involving *g* positions are the next most important (-356kcal/mol or 23% of the total interaction energy) reflecting the proximate position of *g* residues to *a*. The interaction energy between *a-b*, *a-c*, *a-e*, and *a-f* of the two chains is stabilising -342kcal/mol but not as large as between the positions *a* and *d*. The interaction energy between positions *d-b*, *d-c*, *d-e* and *d-f* in both chains is stabilising (-104kcal/mol). The interaction energy between the *a-a*, *a-d*, *d-a* and *d-d* positions are the most important for inter-chain stabilisation. There are no interactions within the cut-off criteria between *b-b*, *b-c*, *b-f*, *c-b*, *c-c*, *c-e*, *c-f*, *c-g*, *e-b*, *f-b*, *f-c*, *f-e* and *f-f* positions and the interaction energy between these positions is therefore taken as zero. There are destabilising

interactions between the *b-e* (10kcal/mol), *b-g* (7kcal/mol), *d-e* (13kcal/mol), *e-d* (56kcal/mol), *e-e* (41kcal/mol), *e-f* (9kcal/mol), *e-g* (19kcal/mol), *f-g* (9kcal/mol), *g-b*

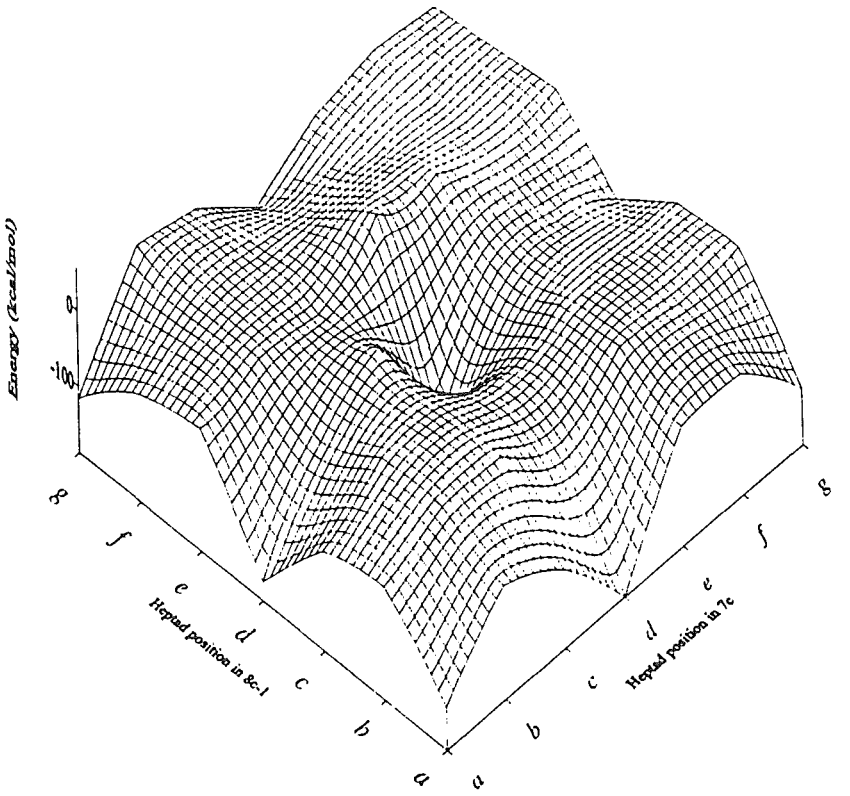


Figure 5.10. 3D energy surface of the interaction vs the heptad position of the two coiled-coil chains.

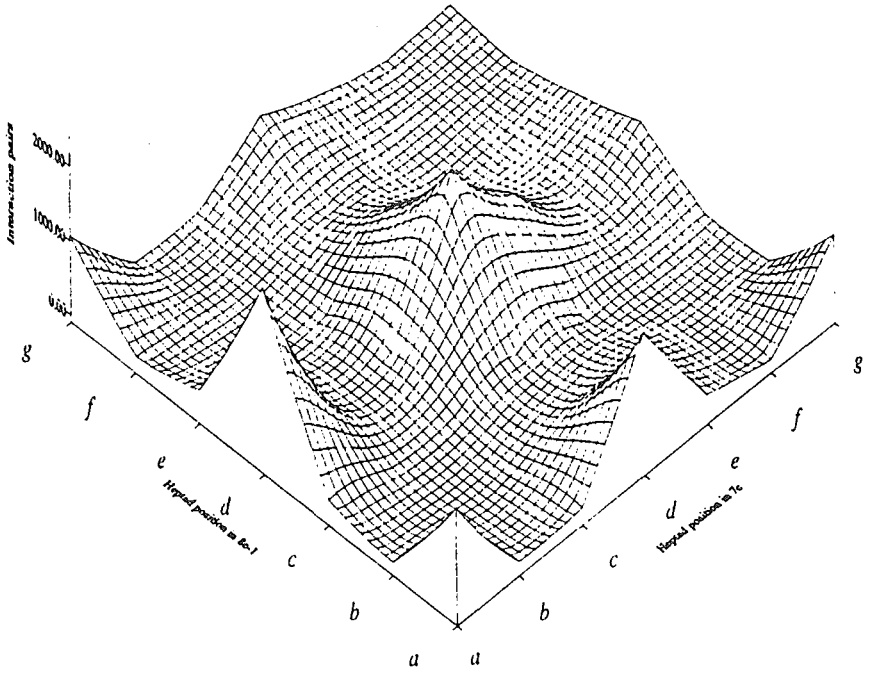


Figure 5.11. 3D surface of atom pairs of interaction vs the heptad position of the two coiled-coil chains.

(19kcal/mol) and *g-f* (8kcal/mol). The interaction energy between heptad residues of linking segments within the cut-off criteria is small (6.5% of the total interaction energy).

Interaction between the *a-a, a-d, d-a* and *d-d* positions of the heptads are the most important. Interaction between the *a-a* (1491), *a-d* (1805), *d-a* (2398) and *d-d* (1952) positions within the cut-off criteria total 7646 or 47% of all interactions. The *g* position of the heptad repeats has 2131 interactions with *a* or *d* positions within the cut-off criteria or 13% of the total number of interactions within the criteria. The interactions involved with position *e* occur primarily with *d* and are not as important as the interactions involving *g*. From the data in Table 5.26, a 3D energy surface of the interactions vs the heptad positions in chain 8c-1 and chain 7c is drawn (Figure 5.10) and a 3D surface of interaction pairs vs the heptad position of chain 8c-1 and 7c is reported in Figure 5.11.

The interaction energy in the heptad positions in both chain 7c and 8c-1 vs the total interaction energy as a percentage for the total interactions is shown in Table 5.27. The interaction between two coiled-coil helices occur primarily at the *a* and *d* positions of heptad repeat. The *e* and *g* positions are also important. Positions other than *a, d, e, g* have few interactions of significance. The linking segments contribute less than 5% of the interactions within cut-off criteria and do not significantly effect the interaction energy of the coiled-coil rod domain section of wool.

Table 5.27. The percentage of the interaction energy corresponding to positions of the heptad repeat.

	Contribution of energy %		Contribution of interaction %	
Position	7c	8c-1	7c	8c-1
a	24.1	38.8	34.9	30.2
b	5.5	4.4	1.9	2.6
c	5.5	5.5	2.1	2.6
d	37.5	29.2	34.8	38.5
e	0.9	1.4	10.1	8.2
f	2.3	2.8	1.4	1.3
g	9.1	10.9	12.1	12.3
linkage	1.9	7.1	2.6	4.6

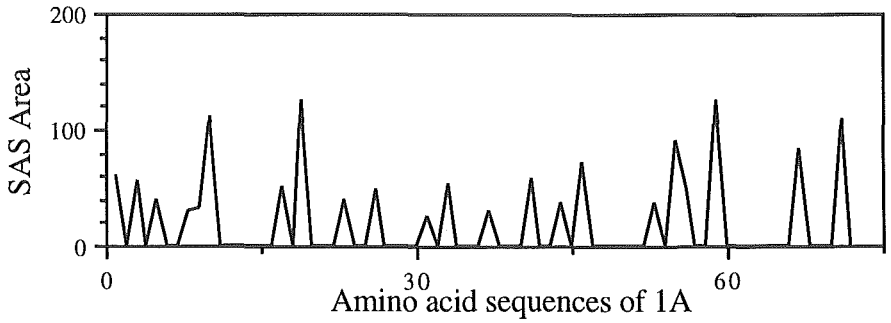
Some 70% of the inter-chain interaction energy is from the *a* and *d* positions (Table 5.27) and 60% of the inter-chain interactions take place between residues at these heptad positions. The interactions involving residues at position *g* contribution 10% to the total interaction energy. The interactions involving position *e* are 10% of the total number of interactions but contribute 1% to the interaction energy. The interactions involving position *b, c, f* can be ignored in the analysis of inter-chain interactions.

5.3.7. The solvent accessible surface (SAS) area and volumes of the rod domain

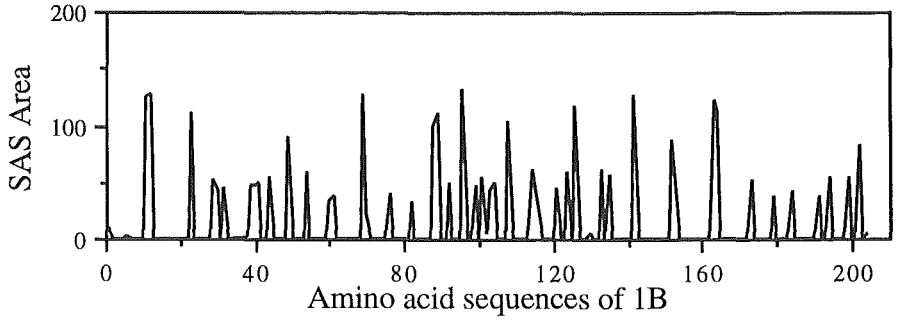
The solvent accessible surface area and volume of the four coiled-coil helical segments of the rod domain have been examined with a molecular probe of radius of 1.4 Å using the program *Gepolt*. The SAS areas of each of the individual amino acid residues are calculated and given in Figure 5.12.

Figure 5.12. The SAS areas of the individual amino acid residues.

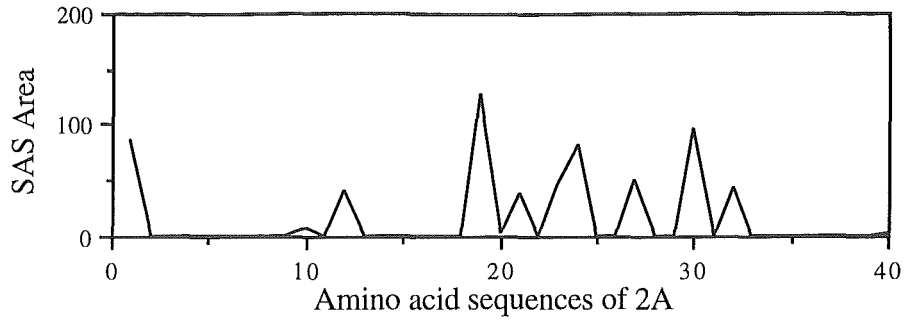
a.



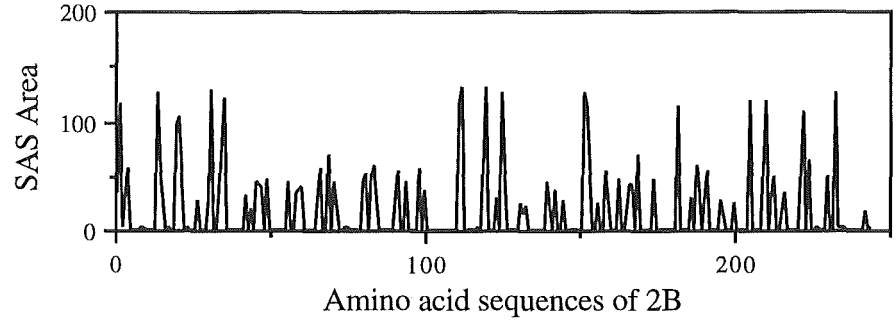
b.



c.



d.



Studies of the structure of globular proteins have shown that in an aqueous environment a polypeptide chain adopts a conformation in which the charged side-chains are exposed to the solvent. The charged and uncharged side-chains therefore exert an important influence on molecular shape.[6] Molecules with a high proportion of charged residues would be expected to assume a shape with a high surface area/volume ratio. Attempts have been made to place such considerations on a quantitative basis. The higher-than-average concentration of charged residues in the coiled-coil α-helical sections of proteins particularly keratin would appear to be related to the high ratio of surface area to volume of the coiled-coil structure. The SAS total areas and volume of each of the four helical coiled-coil segments have been calculated (Table 5.28). The coiled-coil structure is stabilised by the charged residues occupying positions such that they are directed away from the major axis. The ratio of the SAS area/volume are reported in Table 5.28.

Table 5.28. The SAS area, volume and ratio in the helical segments

Segment	SAS (Å^2)	Vol (Å^3)	SAS/VOL
1A	6091	17518	0.348
1B	17924	49121	0.365
2A	3623	9730	0.372

2B	19689	56339	0.349

The surface area/volume ratio may have some relationship to the percentage of exposed charged residues in the helical segments (Table 5.12). The segment 2A of the coiled-coil rod domain has the highest surface area/volume ratio (0.372) compared with that of the segment 1A, 1B and 2B. In segment 2A all the charged residues (100%) are exposed. Segment 2B has 69% of the charged residues exposed, the lowest percentage of the four helical segments. This segment has a lower ratio of surface area/volume than segments 1B and 2A. For segment 1B, the area/volume ratio is 0.358 and the percentage of exposed charged residues (86%) is higher than for segments 2B and lower than for 2A. However, this does not account for why segment 1A where the percentage of exposed charged residues (98%) is higher than 1B and 2B has an area/volume ratio (0.349) lower than 1B and close to 2B. Further investigation is required to address this problem.

5.4. Conclusion

A full atomic model of the coiled-coil rod domain of wool protein has been established by using the MCPB/MCPBA modelling procedure. For the particular knob-hole heptad repeat model investigated the minor helix of the coiled-coil helix, the number of the residues in a turn is between 3.50 and 3.55, the rotation angle per residue 102.9, the pitch is ca. 5.15 Å, the radius is in a range of 2.6 to 2.8 Å. For the coiled-coil helix, the pitch of the rod domain is different in each of the four helical segments and is in the range 124 Å - 190 Å. The radius of the coiled-coil is in the range 5.2 Å - 5.7 Å and varies between the helical segments. An average value of the crossing angle is 23.0°.

The distribution of the residues in the heptad repeats for the model investigated shows 34% of leucine residues in rod domain are located at the *d* position and 25% of leucine residues located at the *a* position. The inter-chain interactions of the coiled-coil helices with the cut-off criteria tas defined are greatest between the *d-a, d-d, a-a* and *a-d* positions.

The assignment of heptad repeats can significantly effect the stability of the structure and further models will need to be generated and investigated for give a more comprehensive understanding of the rod domain of wool. The software reported in this thesis makes this a readily accessible goal. The interaction energy involving charged residues and polar residues are more important than those involving non-polar residues. Excluding aromatic-aromatic interactions the interactions between the non-polar and non-polar residues marginally destabilise the coiled-coil but are positioned between the chains to minimise the destabilising effect they would exhibit in a polar environment. The interactions between the aromatic-aromatic residues contribute -0.06 (kcal/mol)/per interaction. This interaction is stronger than between non-polar non-polar interactions, but weaker than the polar/ionic interactions. The interchain hydrogen bonds of the two helical chains significantly contribute to the inter-chain interaction energy but are limited in number. No cysteine residues were found the chains in the coiled-coil rod domains positioned closely enough to result in formations disulfide bonds between the protein chains.

References

1 Fraser R. D. B., MacRae T. P. (1971) *Nature* **233,** 138-140. Fraser R. D. B., MacRae T. P. Suzuki E. (1976) J. Mol. Biol. **108,** 435-452.

2 Steinert P. M., Idler W. W., Goldman R. D. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77,** 4534-4538.

3 Steinert P. M., Parry D. A. D., Racoosin E. L., Idler W. W., Steven A. C., Trus B. L., Roop D. R. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81,** 5709-5713. Steinert P. M., Idler W. W., Zhou X. M., Johnson L., Parry D. A. D., Steven A. C., Roop D. R. (1985) *Annals of the New York Academy of Sciences* **455,** 451-461.

4 Geisler N., Plessmann U., Weber K. (1982) *Nature* **296,** 488-450.

5 Parry D. A. D., Fraser R. D. B. (1985) *Int. J. Biol. Macromol.* **7,** 203-213.

6 Fraser, R. D. B., MacRae T. P. (1973) *Conformation in Fibrous Proteins and Related Synthetic Polypeptides. Academic Press*, New York.

7 McLachlan A. D. (1978) *J. Mol. Biol.* **124,** 297-304.

8 Weber I. T., Steitz T. A.(1987) *J. Mol. Biol.* **198,** 311-326.

9 Woods E. F., Inglis A. C. (1984) *Int. J. Biol. Macromol.* **6,** 277-283. Parry D. A. D., Crewther W. G., Fraser R. D. B. (1977) *J. Mol. Biol.* **113,** 449-454.

10 Woods E. F., Gruen L. C. (1981) *Aust. J. Biol. Sci.* **34,** 515-526. Geisler N., Weber K. (1982) *EMBO J.* **1,** 1649-1656. Quinlan R. A., Franke W. W. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79,** 3452-3456. Stewart M. Quinlan R. A. Moir R. D. (1989) *J. Cell Biology* **109,** 225-234. Parray D. A. D., Steven A. C., Steinert P. M. (1985) *Bioch&Biophys. Research Comm.* **127,** 1012-1018. Conway J. F., Parry D. A. D. (1988) *Int. J. Biol. Macromol.* **10,** 79-98.

11 Fraser R. D. B., MacRae T. P. (1983) *Biosci. Report* **3,** 517-525. Steinert P. M., Steven A. C., Roop R. D. (1985) *Cell* **42,** 411-419.

12 Steinert P. M. (1990) *J. Biological Chem.* **265,** 8766-8774. Garber A. T., Retief J. D., Dixon G. H. (1989) *EMBO J.* **8,**1727-1734. Fraser R. D. B., MacRae T. P., Suzuki E., Parry D. A. D. (1985) *Int J. Biol. Macromol.* **7,** 258-274.

13 Crewther W. G., Inglis A. S., McKern N. M. (1978) *Biochem. J.* **173,** 365-371. Hanukoglu I., Fuchs E. (1982) *Cell* **31,** 243-254. Hanukoglu I., Fuchs E. (1983) *Cell* **33,** 915-924. Marchuk D., McCrohon S., Fuchs E. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82,** 1609-1613. Quinlan R. A., Cohlberg J. A., Schiller D. L., Hatzfeld M., Franke W. W. (1984) *J. Mol. Biol.* **178,** 365-388. Hatzfeld M., Maier G., Franke W. W. (1987) *J. Mol. Biol.* **197,** 237-255. Milam L. Erickson H. P. (1982) *J. Cell Biology* **94,** 592-596.

14 Crick F. H. C.(1952) *Nature* **170,** 882-883. Crick F. H. C. (1953) *Acta Crystallogr.* **6,** 685-689. Crick F. H. C. (1953) *Acta Crystallogr.* **6,** 689-697.

15 Parry D. A. D., Suzuki E. (1969) *Biopolymers* **7,** 189-197. Parry D. A. D., Suzuki E. (1969) *Biopolymers* **7,** 199-206.

16 McLachlan A. D. (1978) *J. Mol. Biol.* **122,** 493-506.

17 Braun V., Bosch V. (1972) *Proc. Natl. Acad. Sci.U.S.A.* **69,** 970-974.

18 Modelling of the coiled coil structure by molecular dynamics has recently been described by Nilges M., Axel T. B. (1991) *Protein Engineering* **4,** 649-659.

19 Bernstein F. C., Koetzle T. F., Williams E. J. B., Meyer Jr. E. F., Brice M. D., Rodgers J. R., Kennard O., Shimanouchi T., Tasumi M. (1977) *J. Mol. Biol.***112,** 535-542.

20 Phillips Jr G. N., Fillers J. P., Cohen C. (1986) *J. Mol. Biol.* **192,** 111-131.

21 O'Shea E. K., Rutkowski R., Stafford W. F., Kim P. S. (1989) *Science* **245,** 646-648.

22 Banner D. W., Kokkinidis M., Tsernoglou D. (1987) *J. Mol. Biol.* **196,** 657-675.

23 McLachlan A.D., Stewart M. (1975) *J. Mol. Biol.* **98,** 293-304.

24 Levitt M., Greer J. (1977) *J. Mol. Biol.* **114,** 181-293. In a normal straight helical chain the dihedral angles ϕ and ψ of the backbone are distributed in a range $-60^\circ \pm 30^\circ$ and $-40^\circ \pm 20^\circ$ respectively.

25 $x = r_0\cos(\omega_0 t + \varphi_0) + r_1\cos(\omega_1 t + \varphi_1)\cos(\omega_0 t + \varphi_0) + r_1\cos\alpha\sin(\omega_1 t + \varphi_1)\sin(\omega_0 t + \varphi_0)$

$y = - r_0\sin(\omega_0 t + \varphi_0) - r_1\cos(\omega_1 t + \varphi_1)\sin(\omega_0 t + \varphi_0) + r_1\cos\alpha\sin(\omega_1 t + \varphi_1)\cos(\omega_0 t + \varphi_0)$

$z = p(\omega_0 t/2\pi) + z_0 + r_1\sin\alpha\sin(\omega_1 t + \varphi_1)$

Where r_0 = The radius of the major helix (Å). r_1 = The radius of the minor helix (represented by C^α) (Å). φ_0 = The phase angles of the major helix (°). φ_1 = The phase angle of the minor helix (represented by C^α) (Å). ω_0 = The rotation angle of residues in major helix (°). ω_1 = The rotation angle of residues in minor helix (°). M = The number of residues in a major turn. t = 1, 2, 3,.....M (M is the total number of residues in one chain). α = The pitch angle of major helix (°). p = The pitch or the repeat distance on z axis direction in major helix (Å). z_1 = The starting height of the minor helix. N_1 = The number of minor helix turns in a major turn. No = The major helix turns.[14]

26 Dowling L. M., Crewther W. G., Parry D. A. D. (1986) *Biochem. J.* **236,** 705-712.

27 Hatzfeld M., Weber K. (1990) *J. Cell Biology* **110,** 1199-1210.

28 Sparrow L. G., Robinson C. P., McMahon D. T. W., Rubira M. R. (1989) *Biochem. J.* **261,** 1015.

29 The structures of two parallel α-helices are expected to be coiled coil together after energy minimization because the 'knob-hole' interactions of inter-faces of the coiled coils can make the system much stable. We have tested on the protein 2ZTA, two mostly parallel helices with initial pitch of 2000 Å can be automatically coiled to a coiled coil with the pitch 146 Å after energy minimization.

30 The internal coordinates required for generating C^{α} co-ordinates are the distance (d) between two consecutive C^{α} atoms, the angles of three consecutive C^{α} atoms (α) and the torsional angle (τ) of four consecutive C^{α} atoms.

31 X-ray structures of proteins generally have an error in the atomic coordinates of between 0.2 and 0.3 Å. A maximum statistical error of 0.25 Å for every C^{α} position is introduced before the remaining non hydrogen atoms of the backbone from C^{α} positions is carried out and shown to have little effect on the overall model.

32 Chapter 4

33 Jorgensen W. L., Rives T. (1988) *J. Am. Chem. Soc.* **110,** 1657-1666.

34 Macromodel package, Department of Chemistry, Columbia University, New York, NY 10027

35 Still W. C., Tempczyk A., Hawley R. C., Hendrickson T. (1991) *J. Am. Chem. Soc.* **112,** 6127-6129.

36 This is the best general energy minimization method for the protein model larger than 500 atoms. The PRCG commond is chosen in Macromodel energy mode.

37 Kahn P. C. (1989) *Computers Chem.* **13,** 185-189. Kahn P. C. (1989) *Computers Chem.* **13,** 191-195.

38 The torsional angles in a coiled coil structure modelled using Crick equations give very similar values of ϕ and ψ. For an ideal helical motif, the dihedral angles of the backbone, ϕ and ψ, are normally in range of $-60^{\circ} \pm 30^{\circ}$ and $-40^{\circ} \pm 20^{\circ}$ respectively.

39 The program *Gepolt* was from Dr House (personal comminication) and developed by Pascual-Ahuir et al. Pascual-Ahuir J. L., Silla E (1990) *J. Comp. Chem.* **11,** 1048-1060. Silla E., Tunon I., Pascual-Ahuir J. L. (1991) *J. Comp. Chem.* **12,** 1077-1088. Floris F. M., Tomasi J., Pascual-Ahuir J. L. (1991) *J. Comp. Chem.* **12**, 784-791.

40 Lupas A., Dyke M. V., Stock J. (1991) *Science* **252,** 1162-1164.

41 Cohen C., Parry D. A. D. (1990) *Proteins* **7,** 1-15.

42 O'Shea E. K., Klemm J. D., Kim P. S., Alber T. (1991) *Science* **254**, 539-544.

43 Dowling L. M., Crewther W. G., Inglis A. S. (1986) *Biochem. J.* **236,** 705-712.

44 Pauling L., Corey R. B. Branson H. R. (1951) *Proc. Nat. Acad. Sci. U.S* **37,** 205-211. Pauling L., Corey R. B. (1953) *Nature* **171,** 59-61.

45 Fraser R. D. B., MacRae T. P., Miller A. (1964) *Nature* **203,** 1231-1233.

46 Pcmodel package 4.0 version, Serena Software.

47 Marchuk D., McCrohon S., Fuchs E. **1984** *Cell* 39:491-498.

48 Marchuk D., McCrohon S., Fuchs E. **1985** *Proc. Natl. Acad. Sci. U.S.A.* 82:1609-1613.

49 There are four classes of the amino acid residues as following: Acidic amino acids are D (Aspartic acid) and E (Glutamic acid). Basic amino acids are K (Lysine) and R (Arginine). Polar residues are S (serine), T(Threonine), Y(Tyrosine), H (Histidine), C (Cysteine), N (Asparagine), Q (Glutamine) and W (Tryptophan). Non-polar residues are A (Alanine), V (Valine), F (Phenylalanine), P (Proline), M (Methionine), I (Isoleucine), L (Leucine) and G (Glycine).

50 Momany F. A., McGuire R. F., Burgess A. W., Scheraga H. A. (1975) *J. Phy. Chem.* **79,** 2361-2380.

51 Nemethy G., Pottle M. S., Scheraga H. A. **1983** *J. Phys. Chem.* 87:1883-1887.

52 Weiner S. J., Kollman P. A., Case D. A., Singh U. C., Ghio C., Alagona G., Profeta S., Weiner P. (1984) *J. Am. Chem. Soc.* **106,** 765-784. Weiner S. J., Kollman P. A., Nguyen D. T., Case D. A. (1986) *J. Computational Chem.* **7,** 230-252.

53 Baker E. N., Hubbard R. E. (1984) *Prog. Biophys. Molec. Biol.* **44,** 97-179.

54 Chothia C. (1975) *Nature* **254,** 304-308.

55 Ippolito J. A., Alexander R. S., Christianson D. W. (1990) *J. Mol. Biol.* **215,** 457-471. Vedani A., Huhta D. W. (1991) *J. Am. Chem. Soc.* **113,** 5860-5862. Ben-Naim A. (1991) *J. Phys. Chem.* **95,** 1437-1444.

56 Mitchell J. B. O., Price S. L. (1990) *J. Computational Chem.* **11,** 1217-1233.

57 Sippl M. J., Neemethy G., Scheraga H. A. (1984) *J. Phys. Chem.* **88,** 6231-6233.

58 Arnott S., Wonacott A. J. (1966) *J. Mol. Biol.* **21,** 371-383.

59 Standard straight helix means pitch 5.4 Å, residues perturn 3.6, the rotation angle per reisdues is 100°. IUPAC-IUB Commission on Biochemical Nomenclature, Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. (1970) *Biochemistry* **9,** 3471-3479.

60 Burley S. K., Petsko G. A. (1988) *Adv. Protein Chem.* 39, 125-189.

61 Singh J., Thornton J. M. (1990) *J. Mol. Biol.* 211, 595-615.

62 Fraser R. D. B., MacRaeT. P., Sparrow L. G., Parry D. A. D. **1988** *Int. J. Biol. Macromol.* 10:106 -110:

63 Jones D. D., Bernal I. , Frey M. N., Koetzle T. F. (1974) *Acta Crystallogr.* Sect B **30,** 1220-1227.

64 Rogers G. E. *The biology of wool and hair*, Chapman and Hall, London & New York (1988) p152-154.

65 Kauzmann W. (1959) *Adv. Protein Chem.* **14,** 1-63

66 Lee B. (1985) *Biopolymers* **24,** 813-825.

67 Muller N. (1990) *Acc. Chem. Res.* **23,** 23-28.

68 Murphy K. P., Privalov P. L., Gill S. J. (1990) *Science* **247,** 559-561. Doig A. J., Williams D. H. (1991) *J. Mol. Biol.* **217,** 389-398.

69 Scheraga H. A. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 5585-5587.

70 Godzik A., Wesolowski T. (1988) *Biophysical Chemistry* **31,** 29-34.

71 Gengyo-Ando K., Kagawa H. (1991) *J. Mol. Biol.* **219,** 429-441.

72 Hunter C. A., Sanders J. K. M. (1990) *J. Amer. Chem. Soc.* **112,** 5525-5534. Hunter C. A., Singh J., Thorton J. M. (1991) *J. Mol. Biol.* **218,** 837-846.

73 Burley S. K., Petsko G. A. (1988) *Adv. Protein Chem.* **39,** 125-192.

74 Hangauer D. G., Monzingo A. F., Matthews B. W. (1984) *Biochemistry* **23,** 5730-5741.

75 Serrano L., Bycroft M., Fersht A. R. (1991) *J. Mol. Biol.* **218,** 465-475.

Chapter Six

Conclusion

Non-linear regression equations have been developed to establish a relationship between the C^{α} coordinates of a protein and the backbone dihedral angles ϕ and ψ. This relationship has been used to predict absolute values of the dihedral angle ϕ and ψ of a protein backbone knowing only the C^{α} coordinates of the protein. The sign of the dihedral angles ϕ_i and ψ_i of the *i*th amino acid is assigned from a comparison of the C_{i-1}^{α} to C_{i+2}^{α} distance and the C_{i-1}^{α}, C_i^{α}, C_{i+1}^{α}, C_{i+2}^{α} torsional angle with predetermined ranges of these values established to best correlate with sign. For α-helical regions, 98%, 96% and 95%, 91% respectively of the dihedral angles ϕ and ψ fall within $\pm$ 45° and $\pm$ 30° windows of the value in the X-ray structure. For β-sheet regions 96% and 91% fall within $\pm$ 45° window and 88% and 81% within $\pm$ 30° window. The overall accuracy for the prediction of the backbone dihedral angles for the twenty-four proteins is 94% and 91% respectively within a $\pm$ 45° window and 88% and 81% within $\pm$ 30° window. The NLRDT method is most successful in predicting dihedral angles of proteins rich in α-helix and β-sheet. The use of this methodology to build models of coiled coil proteins from C^{α} coordinates therefore offers potential. The secondary structure motif of protein can be assigned by either the *RDA* or *DTC* methods. The latter method is simpler and more accurate. Both methods are somewhat better than previously reported methods and are, like all methods, most successful for the assignment of secondary structure for α-helices and β-sheet. We will report the application of the former method to coiled coil proteins in chapter three.

A MCPB method that allows for the construction of a protein backbone from C^{α} coordinates has been developed. The method requires the C^{α} coordinates to be known but the sequence is not required. The method gives the backbone structures whose

coordinates deviate from the X-ray coordinates by an average of 0.52 Å before energy minimisation and 0.43 Å after energy minimisation with the Opls/Amber force field and GB/SA solvent method comparing favourably with other methods. The computational time to generating the backbone coordinates is cost effective. The method is not demanding in computer time even with inclusion of the energy minimisation. The method is accurate, efficient and robust for the modelling of protein backbones. The modelling technique has the potential when integrated into and used in conjunction with traditional X-ray techniques to speed up structure solution.

A simulated annealing procedure is reported to predict side-chain coordinates. The Monte Carlo Protein Building Anneal method (MCPBA) is a simple method for modelling full atomic structure of proteins starting only with the coordinates of C^{α} atoms and the amino acid sequence. The method is accurate, efficient (40s/residue on an IBM RS/6000) and insensitive to random errors of up to 1 Å in C^{α} position in C^{α} coordinates. For main-chain atoms the r.m.s.d averages as 0.45 Å and for all non-hydrogen atoms r.m.s.d of 1.61 Å. We will subsequently report the use of the MCPBA method to generate the complete atomic coordinates of the coiled-coil structures of rod domain in wool protein. The method is easy to use and does not require a large number of protein data base as required for homology building. The accuracy of the method is at least comparable with methods previously reported but not as accurate as Levitt's SMM which however requires a large protein data base. The MCPBA method is applicable therefore in situations where no suitable data base is available and complements data base homology modelling.

A full atomic model of the coiled coil rod domain of wool protein has been established by using the MCPB/MCPBA modelling procedure. For the minor helix of the coiled coil helix, the number of the residues in a turn is in a range of 3.50 to 3.55 with an average value of 3.52, the rotation angle of per residue 102.9°, the pitch is about 5.15 Å, the radius is in a range of 2.6 to 2.8 Å. For the coiled coil helix, the pitch of the rod domain is different in each of the four helical segments and is in the range 124 Å - 190 Å with an average value of 172 Å. The radius of the coiled coil is in the range 5.2 Å - 5.7 Å and depends on the particular helical segment. The average value of the radius

is 5.56 Å for the rod domain. An average value of the crossing angle is 23.0°. The distribution of the residues in the heptad repeats shows 34% of leucine residues in rod domain are located at the *d* position and 25% of leucine residues located at the *a* position. The inter-chain interactions of the coiled coil helices with cut-off criteria take place between the *d-a*, *d-d*, *a-a* and *a-d* positions. The correct assignment of heptad repeats can significantly effect the stability of the structure. The interaction energy involving charged residues and polar residues are more important than those involving non-polar residues. The interactions between the non-polar and non-polar residues destabilising the coiled coil system but are positioned between the chains to minimise the destabilising effect. The interactions between the aromatic-aromatic residues contribute -0.06 (kcal/mol)/per interaction. This interaction is stronger than between non-polar non-polar interactions, but weaker than the polar/ionic interactions. The hydrogen bonds formed in side-chain between the two helices significantly contribute to the inter-chain interaction energy but are limited in number. No cysteine residues were found the chains in the coiled coil rod domains positioned closely enough to result in formations disulphide bonds between the chains.

Many methods are included in this modelling procedure as detailed in the previous chapters. These have included the assignment of secondary structure from the C^{α} coordinates; regression analysis between dihedral angles and the C^{α} coordinates; transformation of the internal coordinates to Cartesian coordinates; the generation of C^{α} atoms of coiled coil helices with a left-handed supercoil and a right-handed minor helix; the generation of the coordinates of the first residue; the generation of the backbone atom coordinates from C^{α} coordinates; analyses of an averaging technique; simulated annealing for optimising the packing of side-chains; determination of heptad repeats of residues in the coiled coil and energy refinement of the model structures. Results show that the modelled structures are similar to the X-ray structures for both IF proteins and globular proteins.

Further suggested work in this field would be directed to understanding the conformation of the end domains of the non-coiled coil in the microfibril and

simulation of the interation of the coiled-coils in the formation of the tetramer of the microfibril.

Appendix 1

The algorithm of transforming internal coordinates to Cartesian coordinates

The Cartesian coordinates of the atom i (where i = 4, 5, ... n) in the protein are represented by vector P_i and defined by the preceding three atom vectors labelled P_{i-1}, P_{i-2} and P_{i-3}. The internal co-ordinates of the set of three initial redefined atoms, P_{i-1}, P_{i-2} and P_{i-3} are chosen to have bond length d_i from P_i to P_{i-1}, bond angle λ_i defined by P_i, P_{i-1}, P_{i-2}, and dihedral angle χ_i, between P_i, P_{i-1}, P_{i-2}, and P_{i-3}. The range of values for the internal coordinates d_i, λ_i, χ_i, are $[0, \infty]$, $[0, \pi]$, $[-\pi, \pi]$ respectively. A new Cartesian system is created at the coordinates represented by the atoms corresponding to the vectors P_{i-1}, P_{i-2} and P_{i-3} to satisfy the following conditions;

(i) The origin of the new system is first placed on the point corresponding to the vector P_{i-1}.

(ii) x-axis is drawn from P_{i-2} to P_{i-1}.

(iii) The x0y plane of the new coordinate system is the plane defined by the atoms corresponding to the vectors P_{i-1}, P_{i-2}, P_{i-3}. P_i is positioned so that $\lambda_i = 180^\circ$ and $\chi_i = 0^\circ$.

The Cartesian co-ordinates of the atoms corresponding to the vectors P_{i-3}, P_{i-2}, P_{i-1} on the system are $P_{i-1} = (0, 0, 0)$, $P_{i-2} = (-d_{i-1}, 0, 0)$ and $P_{i-3} = (-d_{i-1} + d_{i-2}\cos\lambda_{i-1}, d_{i-1}\sin\lambda_{i-1}, 0)$.

Two operations are utilised during the above transformation (see Figure 1), namely rotation about z-axis by $(\pi-\lambda)$ to satisfy the bond angles P_i, P_{i-1}, P_{i-2}, and rotation about x-axis by χ to satisfy the dihedral angles P_i, P_{i-1}, P_{i-2}, and P_{i-3}.. The Cartesian co-ordinates P_i on the new system can be written as;

$$P_i(x, y, z) = U(\lambda)R(\chi)P_i(d_i, 0, 0).$$

where $U(\lambda)$ and $R(\chi)$ are rotation matrixes about z-axis and x-axis respectively and d_i is the distance (bond length) of P_i to P_{i-1}. The Cartesian co-ordinates of the atom corresponding to the vector P_i on the original system can be determined by transforming the coordinates of the new system into that of original system.

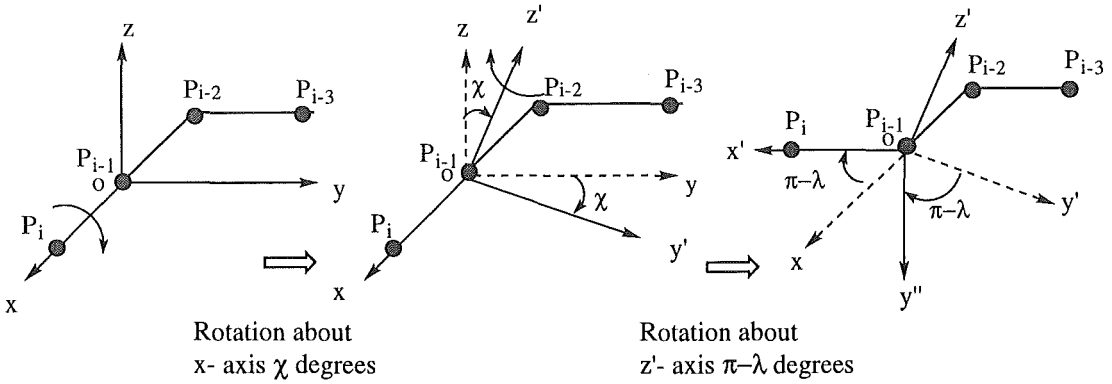


Figure 1. The transformation of internal co-ordinates to Cartesian co-ordinates

---

1 Dudek M. J., Scheraga H. A. *J. Comp. Chemistry* (1990) **11**, 121-151.

Appendix 2.

The determination of the parameters of the helical axis

1. The parameters defining the α-helix

For a simple helix, the parameters which define the helix are the pitch (p), the radius (r), the rotation angle (ω), the rise (d) per residue, the residue number (n) per turn and the helical angle (β). The pitch (p), the rise (d) per residue and residue number (n) per turn have a relation defined by the equations;

$$p = d*n;$$

and $n = 2\pi/\omega.$

The pitch (p), radius (r) and the helical angle (b) has a relation as;

$$\tan\beta = 2\pi r/p$$

For a coiled-coil helix, the parameters of the major helix are pitch (P), radius (r_0), the rotation angle (ω_0) of coiled-coil, the rise (H) per minor turn, the helical angle (β_0) and the number of the residue (N_1) per major turn. The pitch (P) and the number of the residue per major turn (N_1) have a relation defined by the equations;

$$P = H*(2\pi/\omega_0);$$

and $N_1 = P/(H*\cos(\beta_0)).$

The pitch P , radius and helical angle β_0 have a relation as;

$$\tan\beta_o = 2\pi r_o/P$$

2 The method for determining axis of an α-helix

The method for determining the axis of a simple helix of a protein has been developed by P. C. Kahn.[1] The principle of the method is to construct a vector **P1** from the origin to $C^{\alpha}{}_2$. From $C^{\alpha}{}_2$, vectors **A** and **B** are constructed to $C^{\alpha}{}_1$ and $C^{\alpha}{}_3$ respectively. The bisector of the angle defined by **A** and **B** will be perpendicular to the axis of the helix which we wish to define. The vector sum of **A** and **B**, a new vector namely $\mathbf{V_1}$, is perpendicular to the axis of the helix. The new vector $\mathbf{V_1}$ is normalised. The vector $\mathbf{V_2}$ is obtained in the same way as vector $\mathbf{V_1}$ but from $C^{\alpha}{}_2$, $C^{\alpha}{}_3$ and $C^{\alpha}{}_4$. Therefore, the vector $\mathbf{V_i}$ can be obtained in this way from the atoms $C^{\alpha}{}_i$, $C^{\alpha}{}_{i+1}$ and $C^{\alpha}{}_{i+2}$. This process is repeated to $C^{\alpha}{}_{n-1}$. Let $\mathbf{P_{i+1}}$ be the vector from the origin to the $C^{\alpha}{}_{i+1}$ of the $C^{\alpha}{}_i$, $C^{\alpha}{}_{i+1}$ and $C^{\alpha}{}_{i+2}$. Since $\mathbf{V_i}$ and $\mathbf{V_{i+1}}$ are perpendicular to the axis of the helix, the cross product of these two vectors will define the direction of the axis. The cross product of $\mathbf{V_i}$ and $\mathbf{V_{i+1}}$ is made by the equation

$$\sin\omega = (\mathbf{V_i} \times \mathbf{V_{i+1}}) / |\mathbf{V_i}| * |\mathbf{V_{i+1}}| \qquad (1)$$

where ω is the angle between these two vectors $\mathbf{V_i}$ and $\mathbf{V_{i+1}}$.

A vector $\mathbf{H_0}$ is defined to be equal the cross product,

$$\mathbf{H_0} = |\mathbf{V_i}| * |\mathbf{V_{i+1}}| = (\mathbf{V_i} \times \mathbf{V_{i+1}})/\sin\omega.$$

The vector $\mathbf{H_0}$ is Normalised to get a unit vector $\mathbf{h_i}$ by the equation;

$$\mathbf{h_i} = \mathbf{h_i}\, x + \mathbf{h_i}\, y + \mathbf{h_i}\, z \qquad (2)$$

where; $\mathbf{h_i}x$, $\mathbf{h_i}y$ and $\mathbf{h_i}z$ are the component vectors of $\mathbf{h_i}$ in the direction of the x, y and z axis respectively, and

$\mathbf{h}_i x = \mathbf{H}_0 x / |\mathbf{H}_0|$

$\mathbf{h}_i y = \mathbf{H}_0 y / |\mathbf{H}_0|$

$\mathbf{h}_i z = \mathbf{H}_0 z / |\mathbf{H}_0|$

Let r_i be the radius of the helix corresponding the $\mathbf{V}_i$ and d_i the distance between two parallel vectors $\mathbf{V}_i$ and $\mathbf{V}_{i+1}$. $\mathbf{H}_i$ and $\mathbf{H}_{i+1}$ are vectors, from the origin to the intersections of $(r_i)^* \mathbf{V}_i$ and $(r_{i+1})^* \mathbf{V}_{i+1}$ with the helical axis. It follows that

$$\mathbf{H}_i = \mathbf{P}_i + (r_i)^* \mathbf{V}_i \quad (3)$$

$$\mathbf{H}_{i+1} = \mathbf{P}_{i+1} + (r_{i+1})^* \mathbf{V}_{i+1} \quad (4)$$

and

$$\mathbf{H}_{i+1} = d_i \mathbf{h}_i + \mathbf{H}_i \quad (5)$$

To solve these equations, r_i and r_{i+1} must be assumed to be equal. The radius of the axis of the helix at points C^{α}_{i+1} and C^{α}_{i+2} can be obtained;

$$r_i = \{ |d_i \mathbf{h}_i|^2 - |(\mathbf{P}_{i+1} - \mathbf{P}_i)|^2 \} / 2|(\mathbf{P}_{i+1} - \mathbf{P}_i)\mathbf{V}_{i+1}| \quad (6)$$

To solve equation 6, one needs d_i, which is obtained as the projection of $(\mathbf{P}_{i+1} - \mathbf{P}_i)$ on $\mathbf{h}_i$, the axis of the helix, since $\mathbf{V}_i$ and $\mathbf{V}_{i+1}$ are skew lines, ie. lines in parallel planes. Thus d_i is the perpendicular distance between two lines perpendicular to the axis and cutting the helix through the C^{α}s at $\mathbf{P}_i$ and $\mathbf{P}_{i+1}$;

$$d_i = (\mathbf{P}_{i+1} - \mathbf{P}_i)\mathbf{h}_i \quad (7)$$

Substitution of d_i from equation (7) to equation (6) gives r_i, and substitution of r_i into equations (3) and (4) yields two points on the axis, each opposite the central C^{α} of each trio. Successive sets of four C^{α} atoms allows values of the rise per residue, d_i, and the helical radius r_i to be calculated. The angular rotation per residue about the axis, the number of residues per turn, and the pitch can be obtained easily. The cosine of the

angle of rotation (ω_i) is simply $\mathbf{V}_i*\mathbf{V}_{i+1}$. The number of residues per turn (n_i) is therefore obtained by dividing 360° by $\cos^{-1}(\mathbf{V}_i*\mathbf{V}_{i+1})$, and the pitch ($p_i$) by multiplying the number of residues per turn by d_i obtained from equation 7. This method requires the number of residues (m) in the helical segment to be greater than 4. The parameters (radius, rise per residue, the number of residue per turn, the pitch and the helical angle) of the helix are obtained from the average values of r_i, d_i, ω_i, n_i and p_i (i = 1, 2, 3, ... m-1). This completes the description of the helix.

3 The determination of parameters of simple helical segments.

The program *axisc* has been written based on the method described above to calculate the parameters of a single helical segment of a protein. The program reads the coordinates of the X-ray structure in Macromodel format and extracts the C^α coordinates. The α-helix segments can be defined based on the values of the torsional angles of four consecutive C^α atoms. If the torsional angle of four consecutive C^α atoms is less than 90° and larger than 0°, the four consecutive C^α atoms are considered as an α-helical segment. A minimum of four consecutive C^αs are needed to define an α-helix segment. After the α-helix segment is defined, the coordinates of the axis relative to the C^α positions are calculated. To make the axis smooth, the least square method is used to give the best fit to the axis.[2] The coordinates of the axis of an α-helix are used to define the axis of the coiled-coil helix. The program has been tested on the X-ray structures of the proteins 1ROP, 2ZTA and tropomyosin to determine the axis of the helical segments. In addition the theoretical model of MLP has also been examined. The segment 1B of the rod domain in wool protein was generated by the program *supercoil*. (see Table 4).

Table 4. The parameters of the simple α-helix segments in the X-ray structures

Protein	Helical chain	r	d	ω	n	p	m
1ROP	1	2.345	1.353	99.89	3.605	4.872	26

	2	2.360	1.281	98.92	3.643	4.655	23
2ZTA	1	2.331	1.330	99.703	3.612	4.807	28
	2	2.362	1.369	98.568	3.655	5.013	28
MLP	1	2.281	1.427	100.58	3.583	5.136	55
	2	2.280	1.428	100.55	3.585	5.142	55
Tryp.	1	2.415	1.289	99.292	3.628	4.672	281
	2	2.364	1.274	99.312	3.627	4.616	277
Wool.1B	1	2.285	1.441	100.03	3.599	5.187	99
	2	2.285	1.441	100.03	3.599	5.187	99

Note: The amino acid residue is represented by C^{α} atoms. r, the radius (Å); d, the rise per residue (Å); ω, rotation angle (°); n, the number of residues per turn; p, the pitch (Å); m, the number of residues in the helical segment.

The values in Table 4 are experimental average values. The average radii of the amino acid residues (C^{α}) in the α-helix is around 2.3 Å. The average rise of per residues is 1.3 Å. In the protein 1ROP, the average rise per residue in two chains is quite different. The average rotation angle of each residue is about 98° to 99°. This is not the case for the theoretical model of MLP and Wool 1B which is interesting because the rotation angles per residue are less than 100°. This means that there are more than 3.6 residues per turn in the coiled-coil α-helix region of the natural protein. If the number of residues per turn is 3.5, in the supercoil helix of native proteins, which has been the previous assumption,[3,4] the rotation angle should be 102.86°. Such a value has not been observed in the crystal structure of these coiled-coil proteins. The pitch is related to the rise of per residue and residue number per turn. The pitch of α-helix is generally smaller than 5.0 Å in the crystal structures of the coiled-coil helices.

4. Determination of the parameters of a coiled-coil helix.

Appendix 2: The determination of the parameters of the helical axis

The parameters of the coiled-coil helix can also be calculated in the following way. The radius (R_0) of the coiled-coil helix is the distance between the two axis divided by two. The rotation angle (ω_0) is the crossing angle of two vectors: one vector is from ith point of one axis to the relative ith point of the second axis and the second vector is from the (i+1)th point of one axis to the relative (i+1)th point of the second axis. The helix angle (β_0) is the crossing angle of another two vectors: one is from the ith point to the (i+1)th point of one axis and other is from the ith point to the (i+1)th point of second axis. The rise (H) per residue is the length of a vector that from the (i)th point to the (i+1)th point on one axis and then multiplying by $\cos(\beta_0)$. The pitch (P), the number of residues per major turn (N_1) and the length of the segments (S) of the coiled-coil helix are;

$$P = H*(2\pi/\omega_0);$$

$$N_1 = P/[H*\cos(\beta_0)];$$

$$S = M*H*\cos(\beta_0).$$

Where M is the number of residues of the one chain in the coiled-coil helix.

The direction of the two coiled-coil helical chains is defined by the values of the angle β_0 which is generally smaller than 45° for the parallel coiled-coil helical chains. If the value of helical angle β_0 is larger than 45°, the two chains are considered to be anti-parallel. These parameters are listed in Table 5.

Table 5 The parameters of the coiled-coil helix

Code		R_0 (Å)	ω_0	β_0	H (Å)	P (Å)	N1	S (Å)	Direction
1ROP[5]	Exp	4.600	-a	9.750	-	172.5	-	-	Antiparallel
	Calc.	4.430	3.028	9.624	1.514	177.4	118.89	34.3	Antiparallel
2ZTA[6]	Exp	4.650	-	13.50	-	-	-	-	Parallel
	Calc.	4.856	3.742	12.961	1.517	142.2	96.21	41.4	Parallel
MLP[7]	Exp	4.125	-	-	-	186.0	126.0	87.0	Parallel

	Calc.	4.451	2.879	8.843	1.499	185.2	125.03	86.0	Parallel
Tryp. [8,9]	Exp	4.00	-	-	1.500	137.0	-	410.0	Parallel
	Calc.	4.328	3.838	11.185	1.500	138.0	93.79	407.5	Parallel
Wool.1B[10]	Set	4.200	2.800	8.071	1.514	186.0	126.00	-	Parallel
	Calc.	4.263	2.851	8.277	1.492	186.6	126.27	146.3	Parallel

a No value was reported. Note; R_0 radius of coiled-coil helix (Å); ω_0 rotation angle of the axis of minor helix; β_0 the helical crossing angle of coiled-coil helix; H, the rise per minor turn (Å); P the pitch of the coiled-coil helix (Å); N_1 the number of residues in one major turn; S the length of the coiled-coil segments (Å). Exp, means these parameters are taken from the previous reports in references as given. Calc, means these parameters are calculated by program *axisc*.

From the calculations, the radius of the coiled-coil helix segments in the X-ray structures is found to be in the range 4.0 Å to 5.0 Å. The pitch of the coiled-coil helixes varies from 137 Å to 177 Å depending on the proteins. Only 1ROP is an anti-parallel coiled-coil helix observed. 2ZTA and tropomyosin are parallel coiled-coil helices.

---

1 Kahn P. C. *Computers Chem.* (1989) **13,** 185-189.

2 Kahn P. C. *Computers Chem.* (1989) **13,** 191-195.

3 Parry D. A. D., Suzuki E. *Biopolymers* (1969) **7,** 189-198.

4 Parry D. A. D., Suzuki E. *Biopolymers* (1969) **7,** 199-206.

5 Banner D. W., Kokkinidis M., Tsernoglou D. *J. Mol. Biol.* (1987) **196,** 657-675.

6 O'Shea E. K., Klemm J. D., Kim P. S., Alber T. *Science* (1991) **254,** 539.

7 Mclachlan A. D. *J. Mol. Biol.* (1978) **122,** 493-506.

8 Phillips G. N., Fillers J. P., Cohen C. *J.Mol. Biol.* (1986) **192,** 111-131.

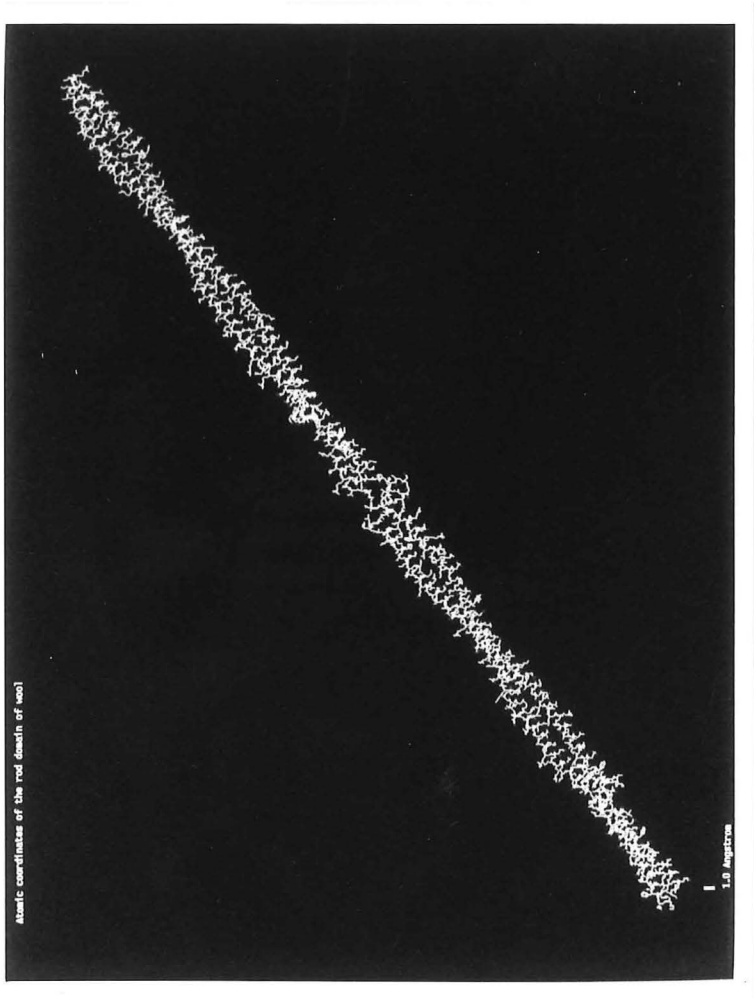
9 Maclahlan A. D., Stewart M. *J. Mol. Biol.* (1975) **98,** 293-304.

10 Wool 1B was generated by program *supercoil*. The parameters of the coiled-coil helix are taken as pitch = 186 Å and radius = 4.2 Å.

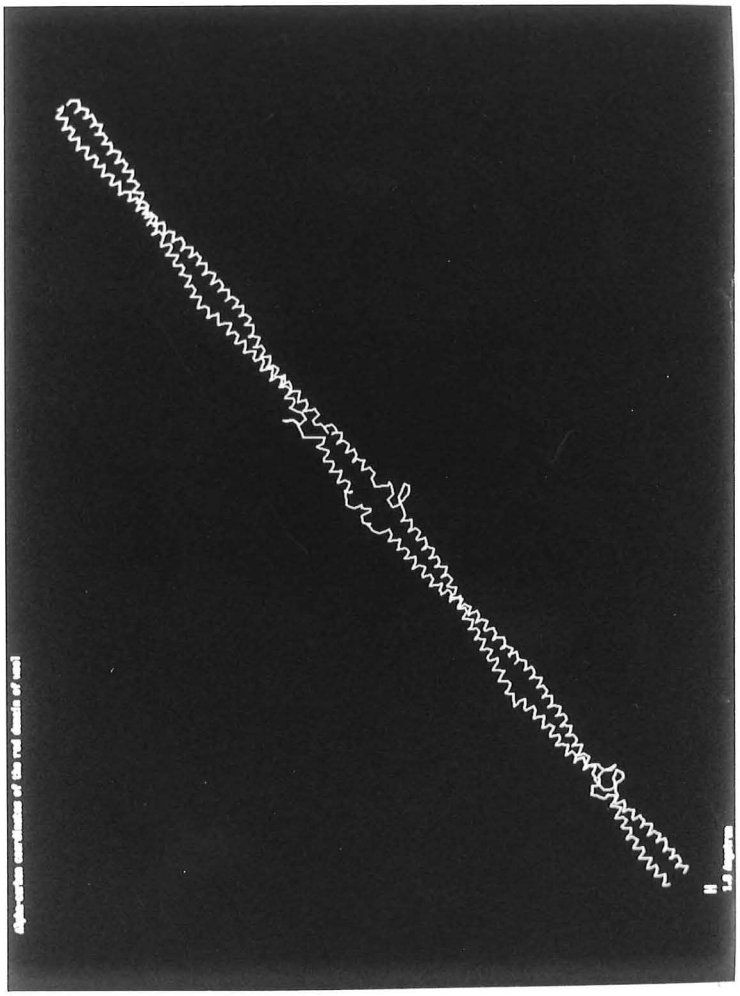
Appendix 3

The pictures of the model of the coiled coil rod domain

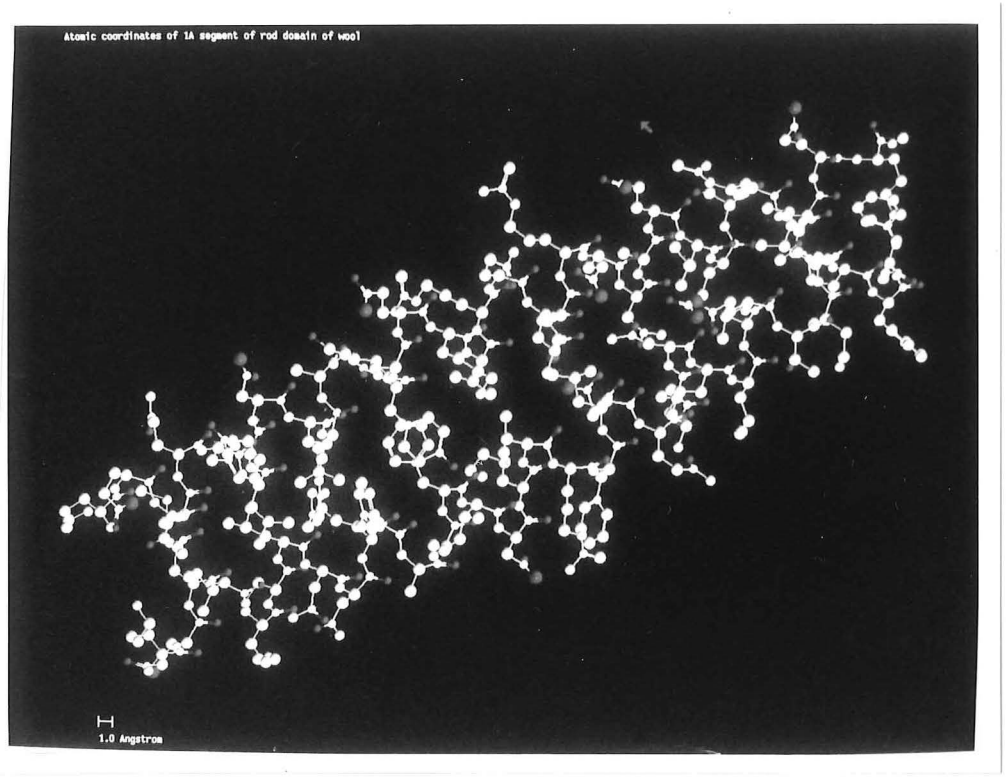
1. Full atom model of the coiled coil rod domain in wool fibre



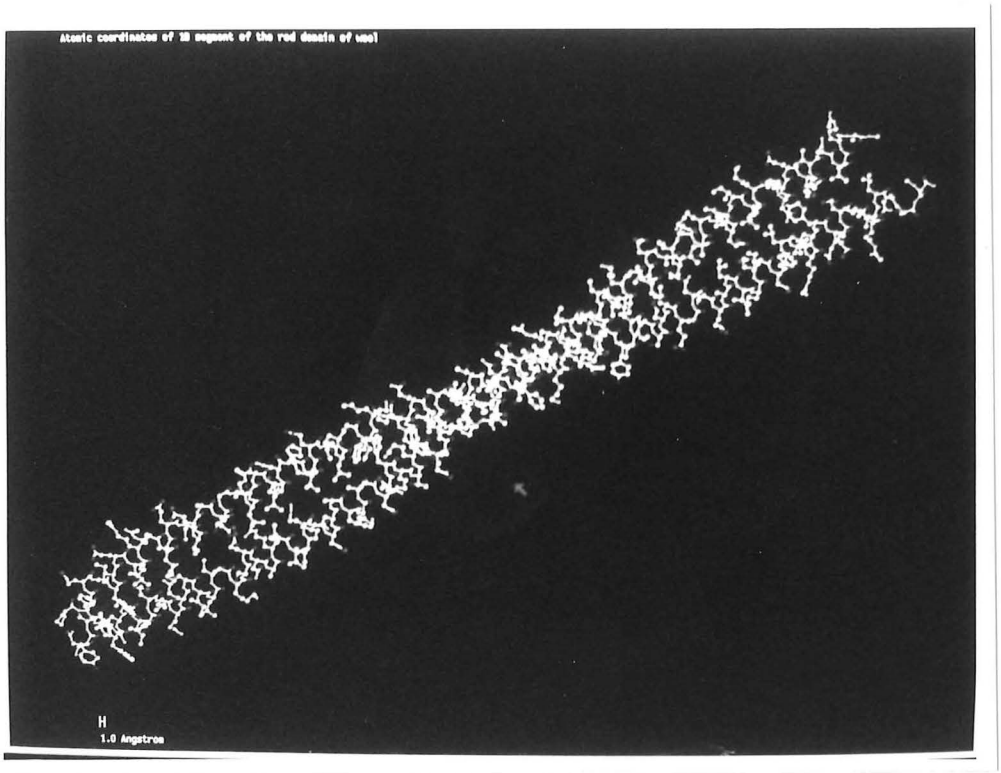
2. The C^{α} model of the coiled coil rod domain of wool fibre



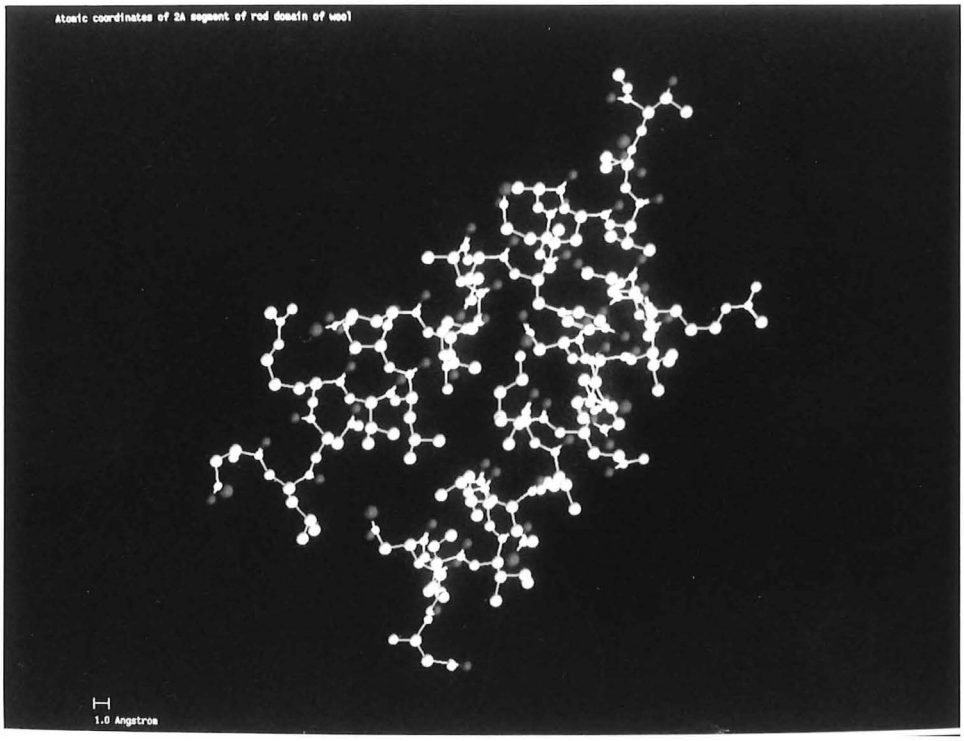
3. Full atom model of the helical segment 1A of the coiled coil rod domain



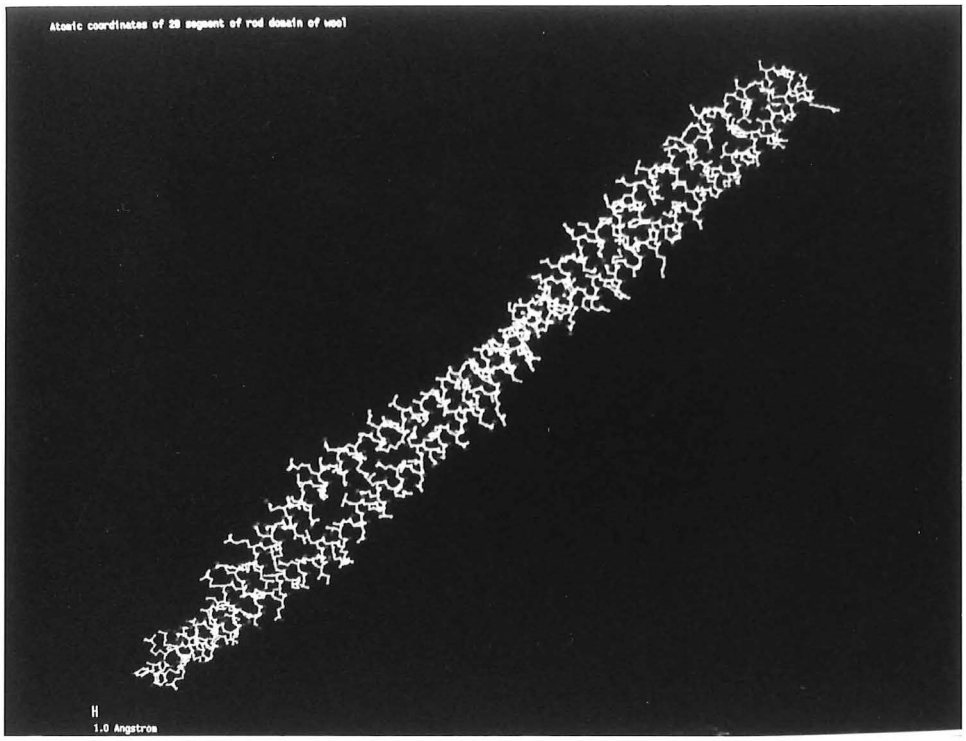
4. Full atom model of the helical segment 1B of the coiled coil rod domain



5. Full atom model of the helical segment 2A of the coiled coil rod domain



6. Full atom model of the helical segment 2B of the coiled coil rod domain



Acknowledgements

Five years have past since I arrived in New Zealand on October 1990. I would like to extend a special thankyou to my supervisor Professor James M. Coxon who gave me the opportunity to undertake the research in New Zealand and for providing a comfortable and stimulating research surrounding. I am very grateful for the help from Professor James M. Coxon and his family during the course of my study in New Zealand.

I would like to thank Dr John Mckinnon and WRONZ for financial support, Dr Gill H. Worth for her helpful suggestions especially in the beginning of this research, Dr Quentin McDonald for his helpful advice during the course of my study, Dr Andrew Burritt for his helpful discussions and Dr Edward Coxon for his friendly help in every area.

I would also like to thank Dr Peter Alexander for his helpful discussion on statistical analysis, Dr Robert Maclagan and Dr House for friendly support and for the program 'Gepol' to calculate the surface and volume of the protein models, Dr Peter Harland for the program 'winsurf' to draw 3D diagrams, and Dr Colin Freeman for his encouragement.

I am also indebted to the members of my research group, Aaron Thorpe, Andrew Cameron, and Karen Lundie for their help. Thanks also go to Mr John Davis for his successful management of the computer system and Mr Alistair Duff for making photos and slides. Finally, I would like to thank Ms Wendy March and Ms Helen Nisbet for their help.